Optimal Prediction, Model Discovery and Self-Organization

Cosma Shalizi
Statistics Department, Carnegie Mellon University

Complex Systems Summer School 2006, Beijing

Part 1: Mostly optimal prediction

complexity of prediction ☢ meaning of “optimal prediction” ☢ causal states ☢ properties of causal states ☢ optimality of causal states

Part 2: Mostly model discovery

the CSSR algorithm ☣ its time complexity ☣ its convergence ☣ hidden state models ☣ other tools for finding such ☣ some synthetic examples ☣ some real data

Part 3: Mostly self-organization

statistical complexity ☢ spatio-temporal systems ☢ self-organization ☢ finding coherent structures ☢ efficiency of prediction ☢ emergence

Mostly optimal prediction

Complexity of prediction

Induction - how long do we need to observe it to learn a good model?

Learning theory (VC dimension etc.); depends on the models we use

Estimation - how much information would we need to make a forecast, if we had the right model?

Calculation - how involved is it to actually calculate the forecast?

System calculates its future at 1 second/second (but see C. Moore, J. Machta, &c.)

Notation

Upper-case letters are random variables, lower-case letters their realizations

Stochastic process: $X_1, X_2, \ldots X_t, \ldots$

Past up to and including time t: X^-_t

Future going forward from t: X^+_t

Making a prediction

Look at X^-_t

Make a guess about X^+_t

Most general guess: distribution of X^+_t

We only attend to some aspects of X^-_t

mean, variance, phases of three largest Fourier modes, ...

so our guess is a *function* or *statistic* of X^-_t

what's a good statistic?

Predictive sufficiency

For any statistic σ

$$I[X^+_t;X^-_t] \geq I[X^+_t;\sigma(X^-_t)]$$

σ is *sufficient* if

$$I[X^+_t;X^-_t] = I[X^+_t;\sigma(X^-_t)]$$

If σ is sufficient, then we only need to know it to minimize any loss function (Blackwell-Girshick)

- σ is sufficient if
- $I[X_{t+1};X_t^-] = I[X_{t+1};\sigma(X_t^-)]$ (one-step ahead)
- $\sigma(x^-_{t+1}) = T(\sigma(x^-_t),x_{t+1})$ for some T (recursion)

Causal states

(Crutchfield & Young 1989)

past *a* and past *b* equivalent iff

$$\Pr(X_t^+ | X_t^- = a) = \Pr(X_t^+ | X_t^- = b)$$

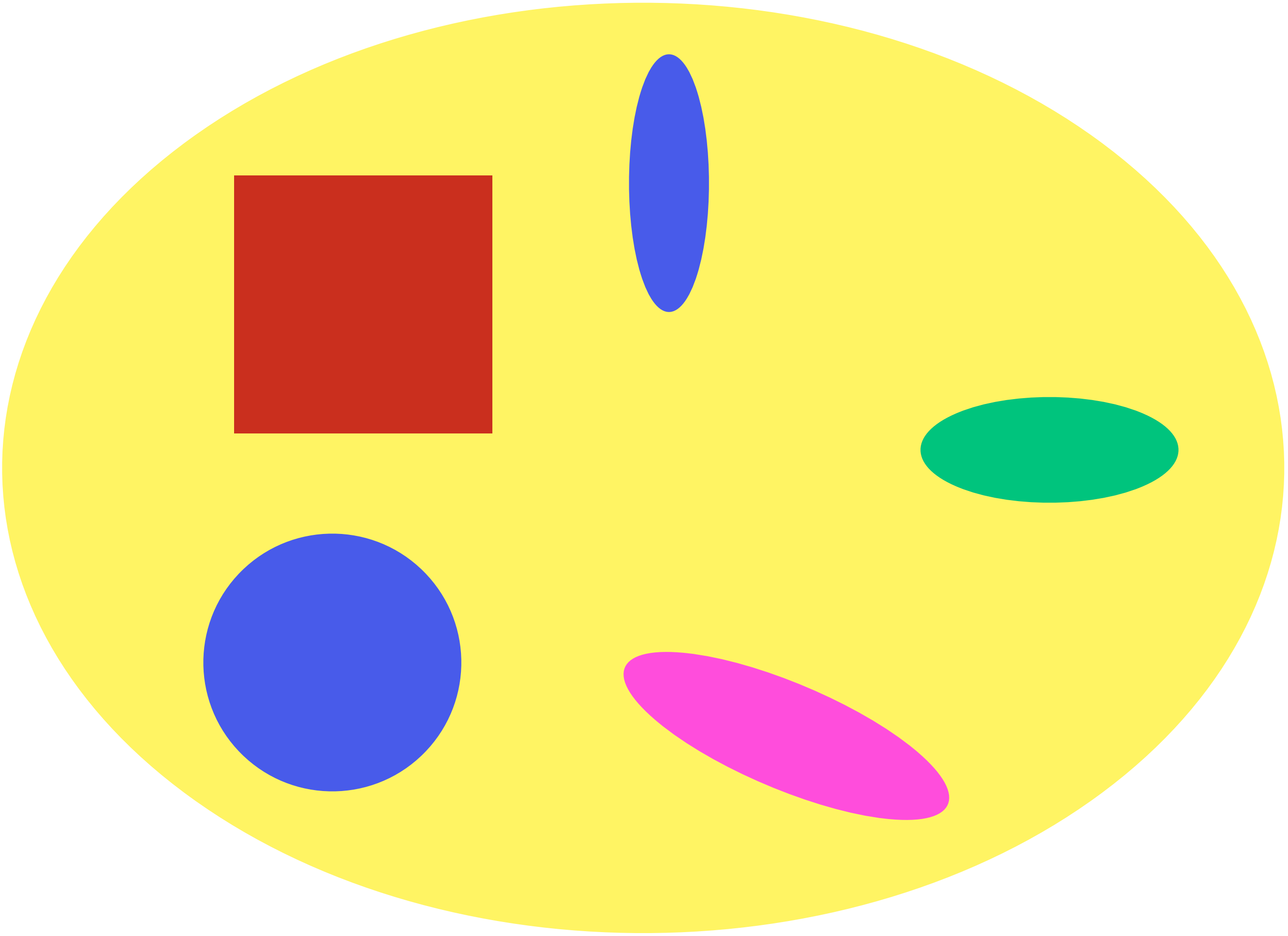
$[a]$ = all pasts equivalent to *a*

Statistic ("causal state"):

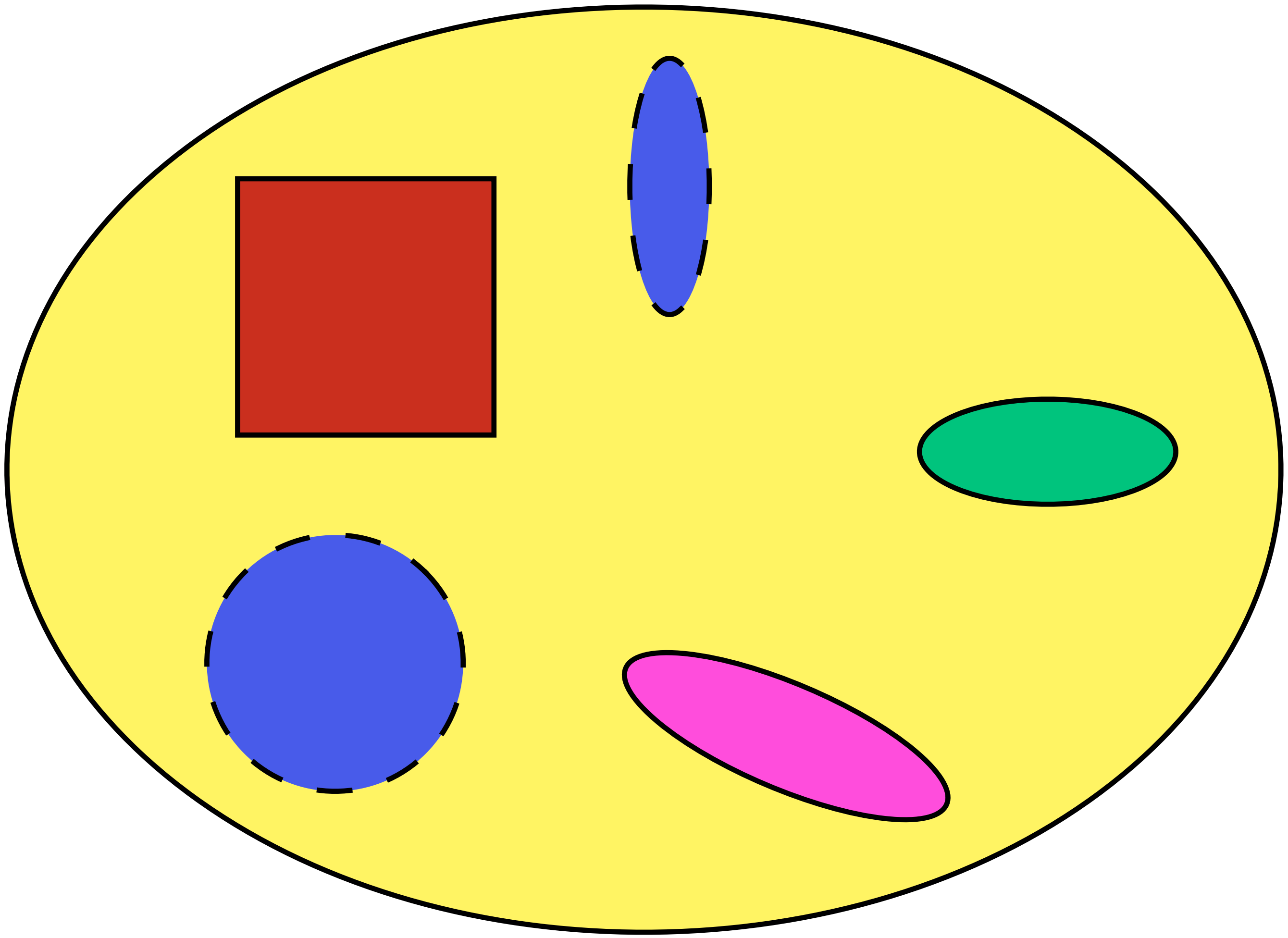
$$\epsilon(x_t^-) = [x_t^-] = s_t$$

State ≡ conditional distribution ≡ histories

IID = 1 state, periodic = p states, ...



Histories and their conditional distributions



Partitioning histories into causal states

History

* “Statistical relevance basis” (Salmon 1971)
“Measure-theoretic prediction process” (Knight 1975)
* “Forecasting / true measure complexity” (Grassberger 1986)
“ϵ-machine” / “causal state model” (Crutchfield & Young 1989)
“Observable operator model” (Jaeger 1999)
“Predictive state representation” (Littman, Sutton & Singh 2002)

Markov properties

(Shalizi & Crutchfield 2001)

Future is independent of past given state

$$X_t^+ \perp X_t^- \mid S_t$$

$\therefore$ Recursive transitions for states

$$\epsilon(x^-_{t+1}) = T(\epsilon(x_t^-), x_{t+1})$$

$\therefore$ States are Markovian

$$S_{t+1} \perp S_{t-1} \mid S_t$$

Optimality properties

(Shalizi & Crutchfield 2001)

Sufficiency:

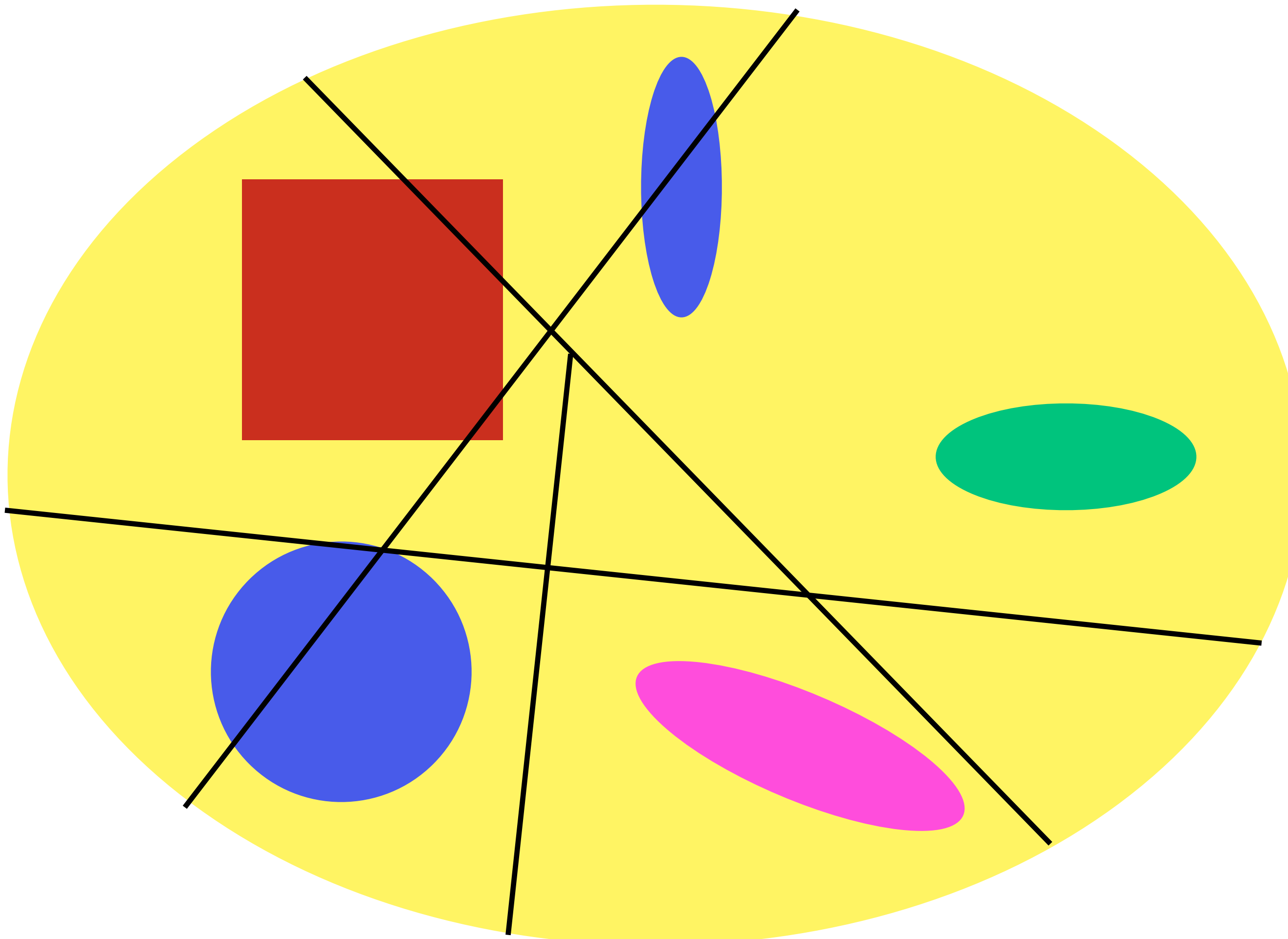
$$I[X_t^+; X_t^-] = I[X_t^+; \epsilon(X_t^-)]$$

because

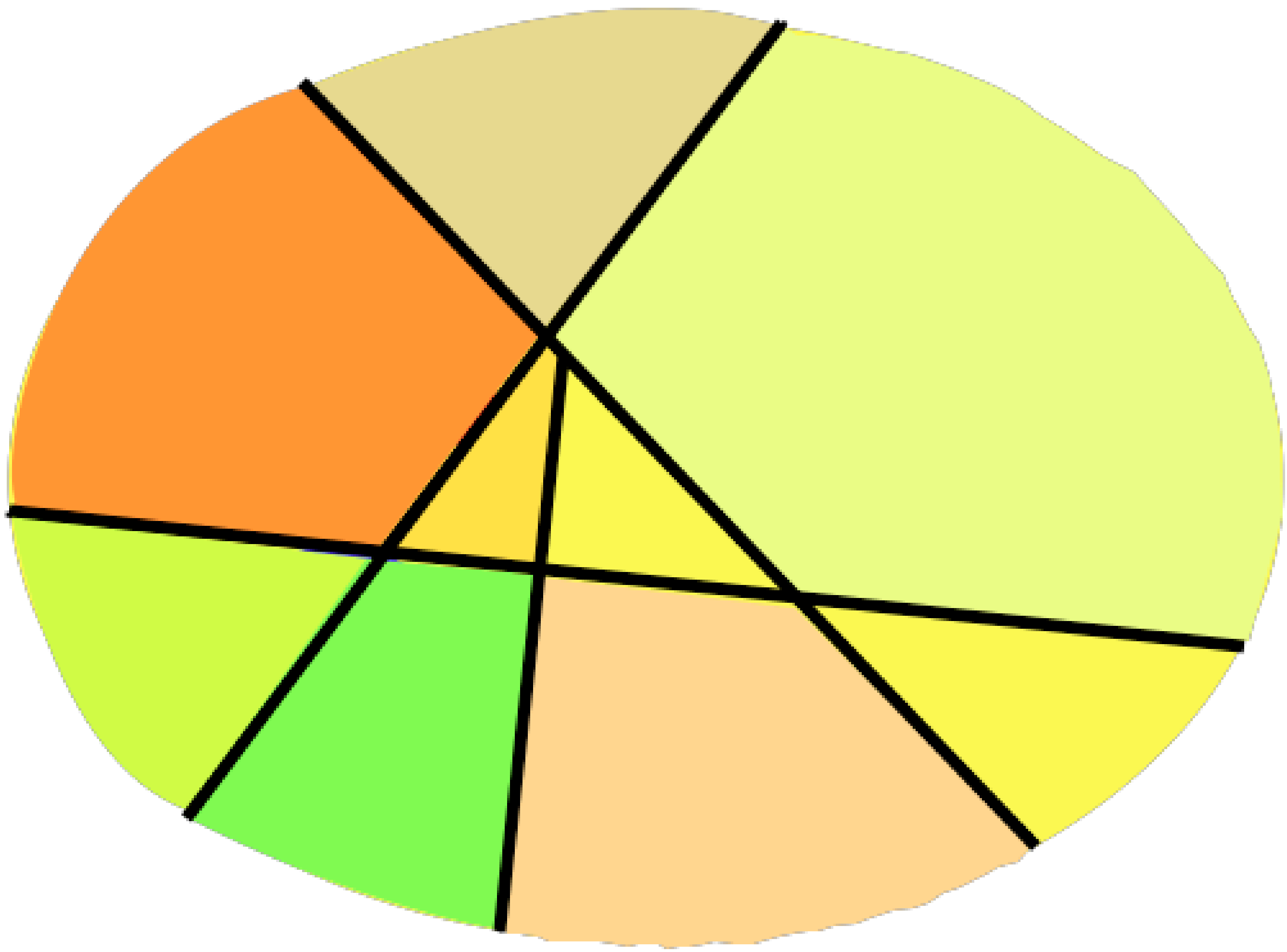
$$\Pr(X_t^+ | S_t = \epsilon(x_t^-))$$

$$= \int_{y \in [x_t^-]} \Pr(X_t^+ | X_t^- = y) \Pr(X_t^- = y | S_t = \epsilon(x_t^-)) \, dy$$

$$= \Pr(X_t^+ | X_t^- = x_t^-)$$



A non-sufficient partition



Effect of insufficiency on predictive distributions

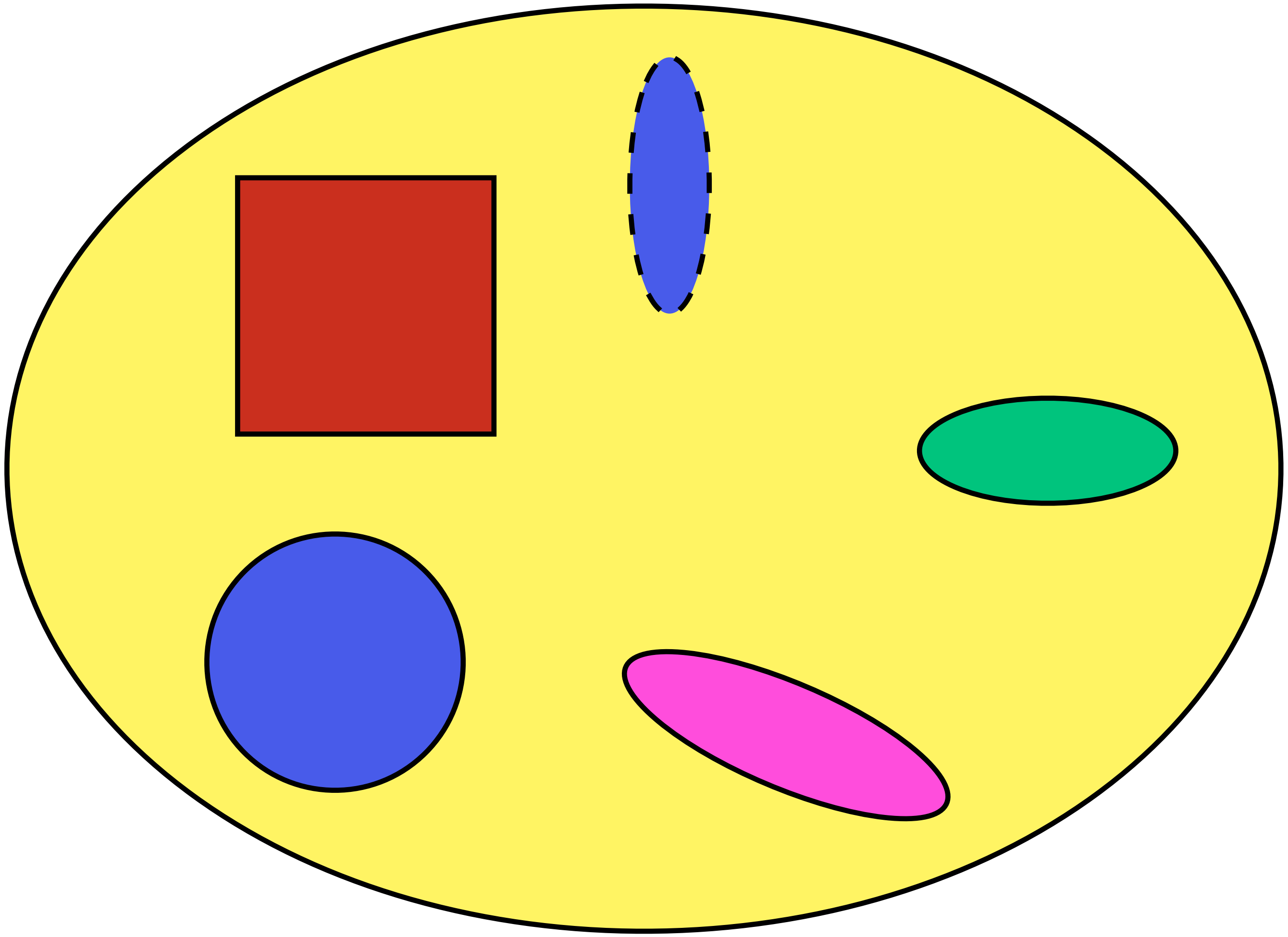
Minimality

Can compute $\epsilon(X_t^-)$ from any other sufficient statistic: for any sufficient η there exists a function g such that

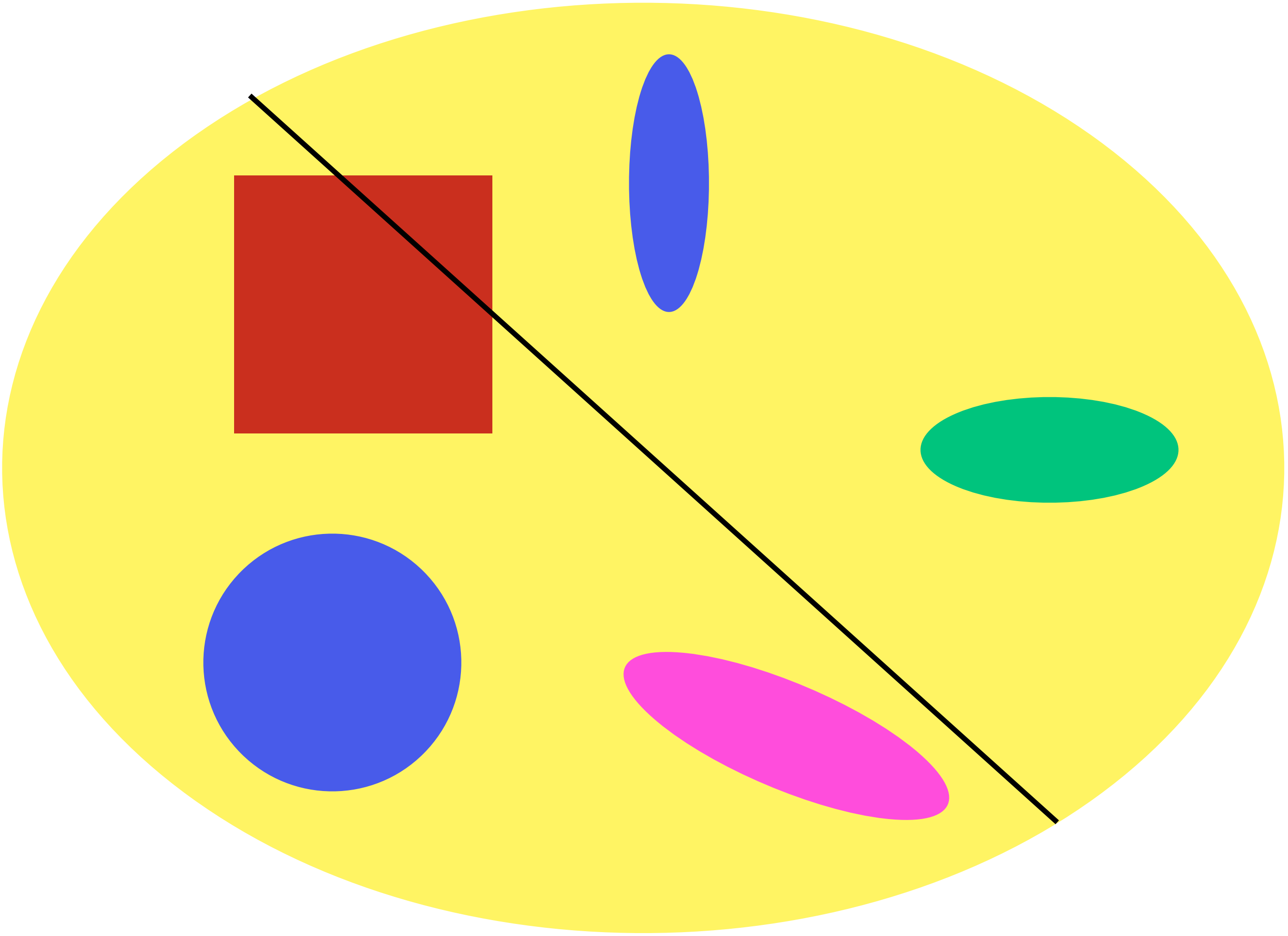
$$\epsilon(X_t^-) = g(\eta(X_t^-))$$

Therefore, if η is sufficient,

$$I[\epsilon(X_t^-);X_t^-] \leq I[\eta(X_t^-);X_t^-]$$



Sufficient, but not minimal



Coarser than the causal states, but not sufficient

Statistical complexity

$C \equiv I[\epsilon(X_t^-);X_t^-]$ is the *statistical* or *forecasting complexity* of the process

$=H[\epsilon(X_t^-)]$

= amount of relevant information stored in the state

= average-case algorithmic sophistication

= log(period) for periodic processes

= log(geometric mean(recurrence time)) for stationary processes

= information about microstate in macroscopic observables (sometimes)

Uniqueness

There is no other minimal sufficient statistic

If η is minimal, there there is an h such that

$$\eta = h(\epsilon)$$

but $\epsilon = g(\eta)$ so

$$g(h(\epsilon)) = \epsilon$$
$$h(g(\eta)) = \eta$$

$g = h^{-1}$ and ϵ and η partition histories in the same way

Minimal stochasticity

If R_t is also sufficient, then

$$H[R_{t+1}|R_t] \geq H[S_{t+1}|S_t]$$

Meaning: the causal states are the closest we get to a deterministic predictive model

Mostly model discovery

CSSR

Causal State Splitting Reconstruction

(Shalizi & Klinkner 2004)

Key observation:

Recursion + next-step predictive sufficiency

$\Rightarrow$ general predictive sufficiency

Get next-step distribution right

Then make states recursive

Assume discrete observations & time, conditionally stationary

http://bactra.org/CSSR

Start with one state, as if IID (history length = 0)

For each state, see if adding one symbol to histories in state makes a difference

If no, go to the next state

If yes, does the new distribution match an existing state?

- *Yes:* move extended history to that state
- *No:* move extended history into a new state

Stop when maximum history length reached

Recursion

Do all the histories in a state make the same transition on the same symbol?
If not, split the state
Keep checking until no state needs to be split

Time Complexity

One pass through data

n data points, k symbols, maximum history length L

Everything-goes-wrong upper bound

$$O(n) + O(k^{2L+1})$$

L can be $\approx \log(n)/(\text{entropy rate})$ [Marton and Shields]

Convergence

S = true causal state structure

S(n) = structure inferred from n data-points

Assume: finite # of states, every state has a finite history, using long enough histories

$$\mathrm{Prob}(S(n) \neq S) \to 0$$

D = true distribution, D(n) = inferred

Error (L^1) scales like independent samples

$$E[|D(n) - D|] = O(n^{-1/2})$$

The Competition: Hidden State Models

What we can see is ugly (non-Markovian, non-stationary, etc.)

Hidden state: what we *can't* see is nice

Usually: guess structure, see if it works

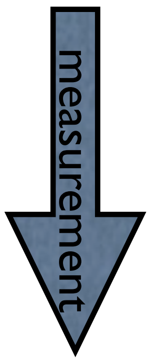
- EM algorithm for parameters + states; Bayesian updating for state estimation
- Mis-specification; complexity

State-space reconstruction

- Entirely data-driven (Ruelle; Farmer, Packard, Crutchfield, Shaw; Takens)
- No EM or Bayes needed
- No good with stochastic dynamics

State

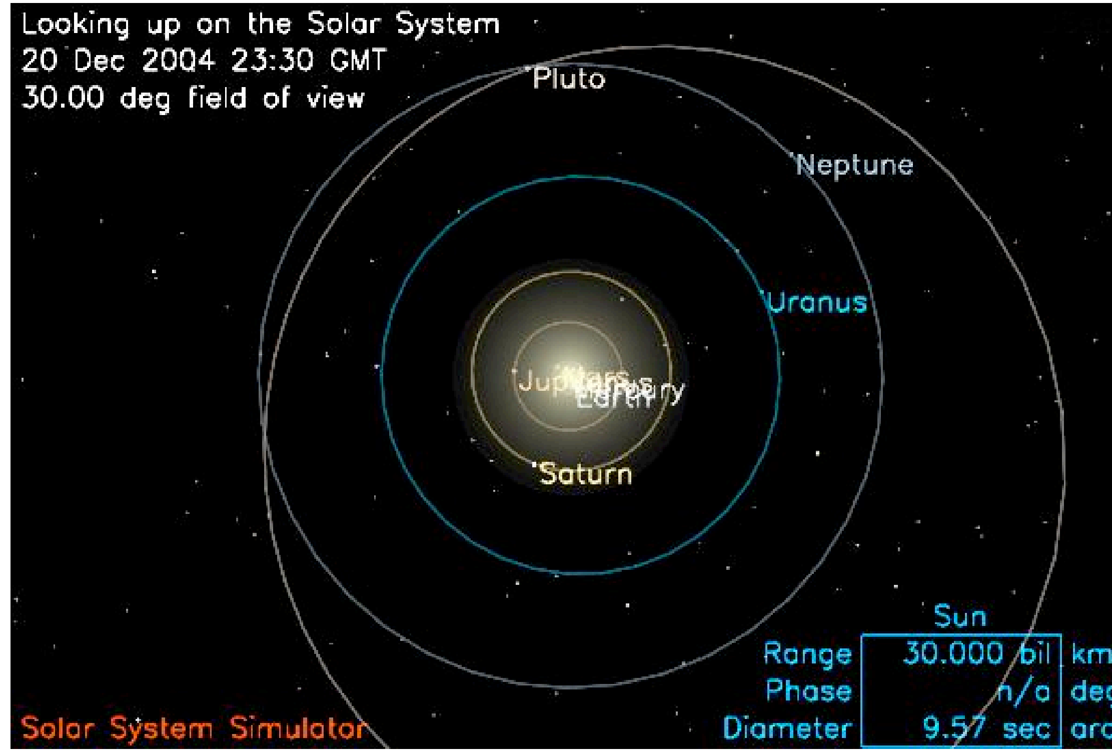
planets in space
54 dimensions



background light
resolution
instrument noise
atmospheric distortion
anatomical distortion
physiological noise
caffeination level
etc.

Observables

lights in sky
14 dimensions



Hidden Markov models

Unobserved states S_t form a Markov process

Observation X_t = random function of S_t

Usually assume $S_{t+1} \perp X_t \mid S_t$ - not here!

Correspond to automata

Variable-length Markov models

(Ron, Singer and Tishby; Buhlmann and Wyner; Kennel and Mees)

a.k.a. Context Trees, Probabilistic Suffix Trees, ...

Split states so that state ≡ suffix

Automatically recursive

VLMM $\subset$ CSSR

CSSR $\not\subset$ VLMM

EM Algorithm + CV

Pick HMM architectures, fit with Expectation-Maximization (Baum-Welch), use cross-validation to select model

Standard heuristic start: fully-connected HMM, with equiprobable state transitions

Selective (not constructive); needs multiple optimizations

Examples

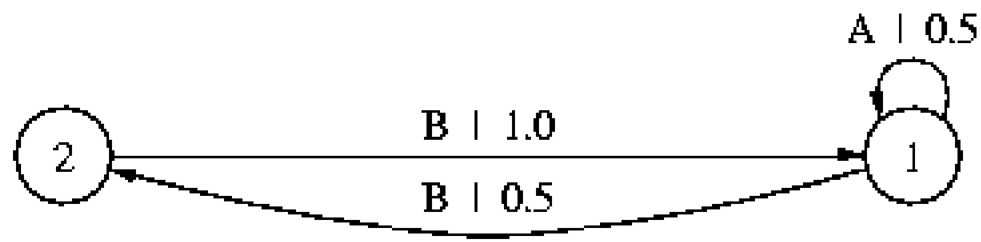
The even process (very trivial)
Foulkes process (trivial)
Model neuron (perhaps not trivial)
Real neuron

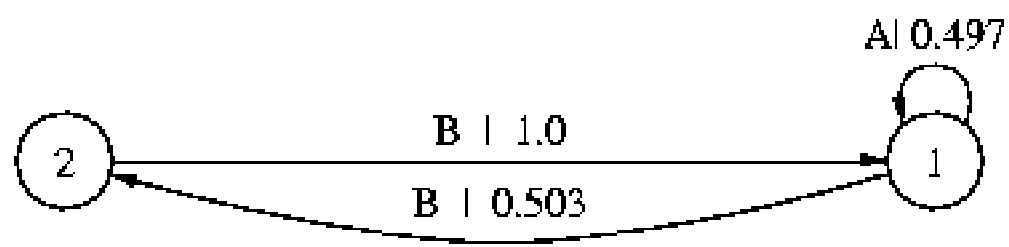
The even process

Language: blocks of A's, any length, separated by blocks of B's, even length

Infinite-range correlation

Reconstructed with history length 3





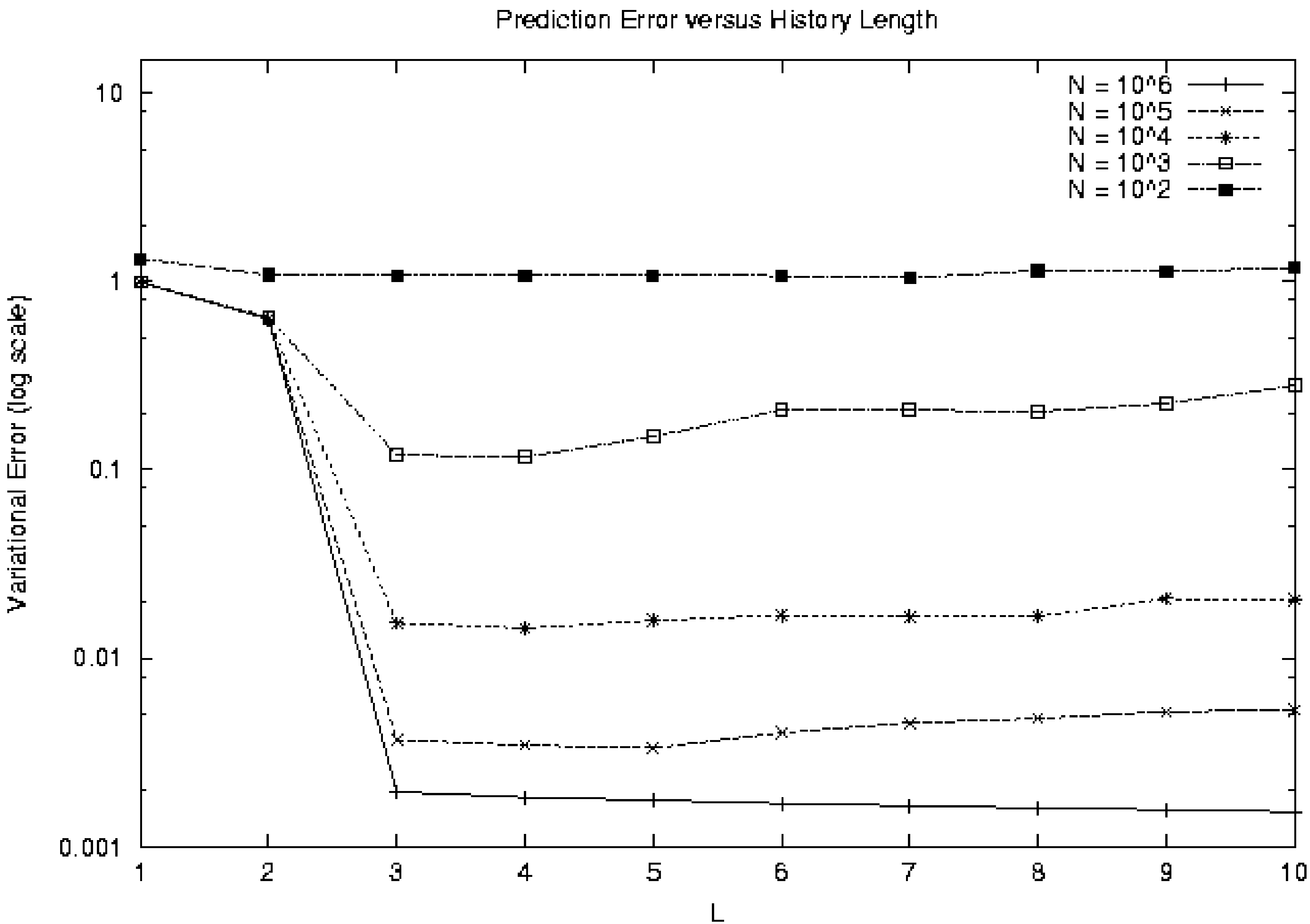
with 10,000 symbols

States as classes of histories:

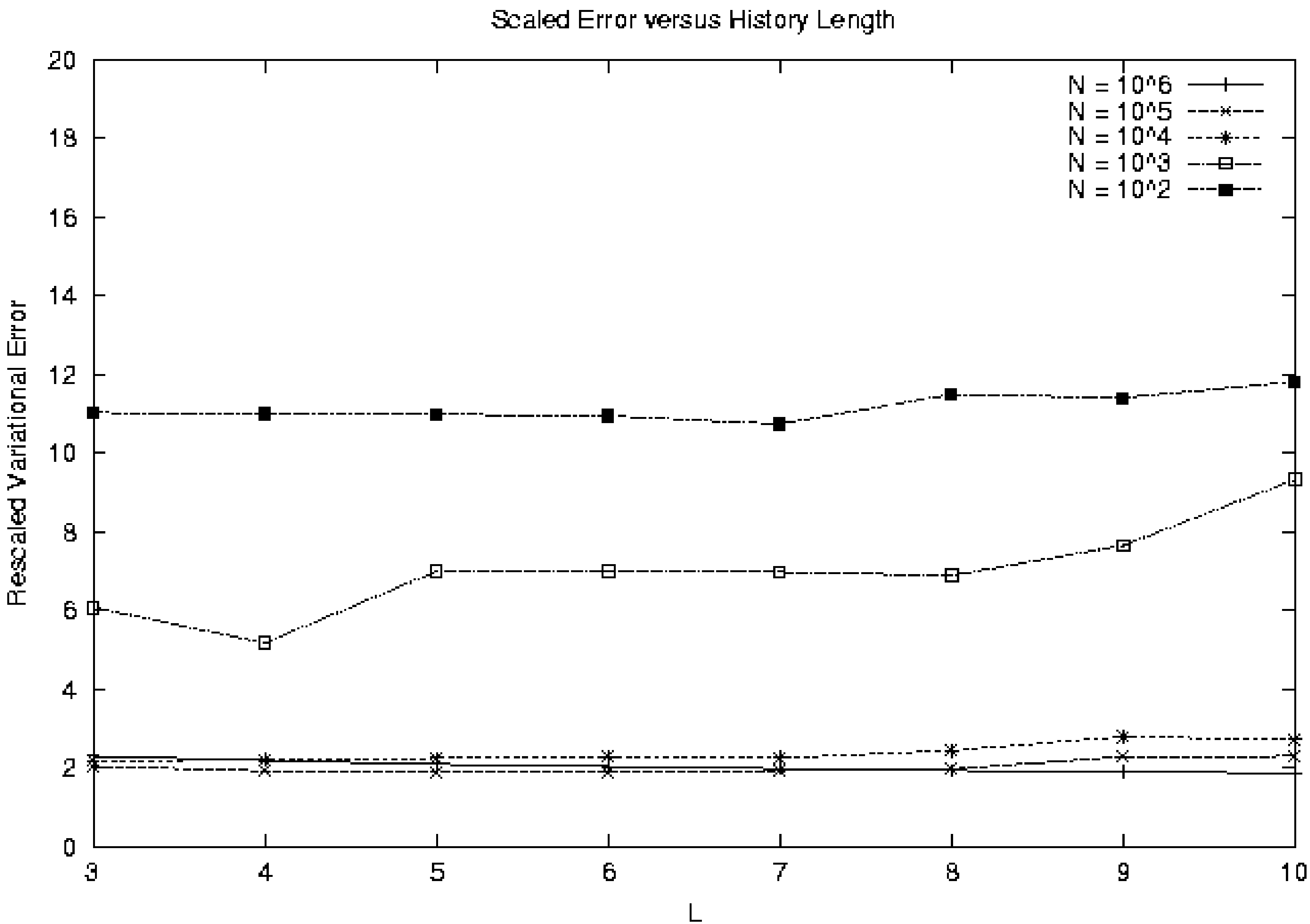
State 1 = *A, *ABB, *ABBBB, etc.
State 2 = *AB, *ABBB, *ABBBBB, etc.

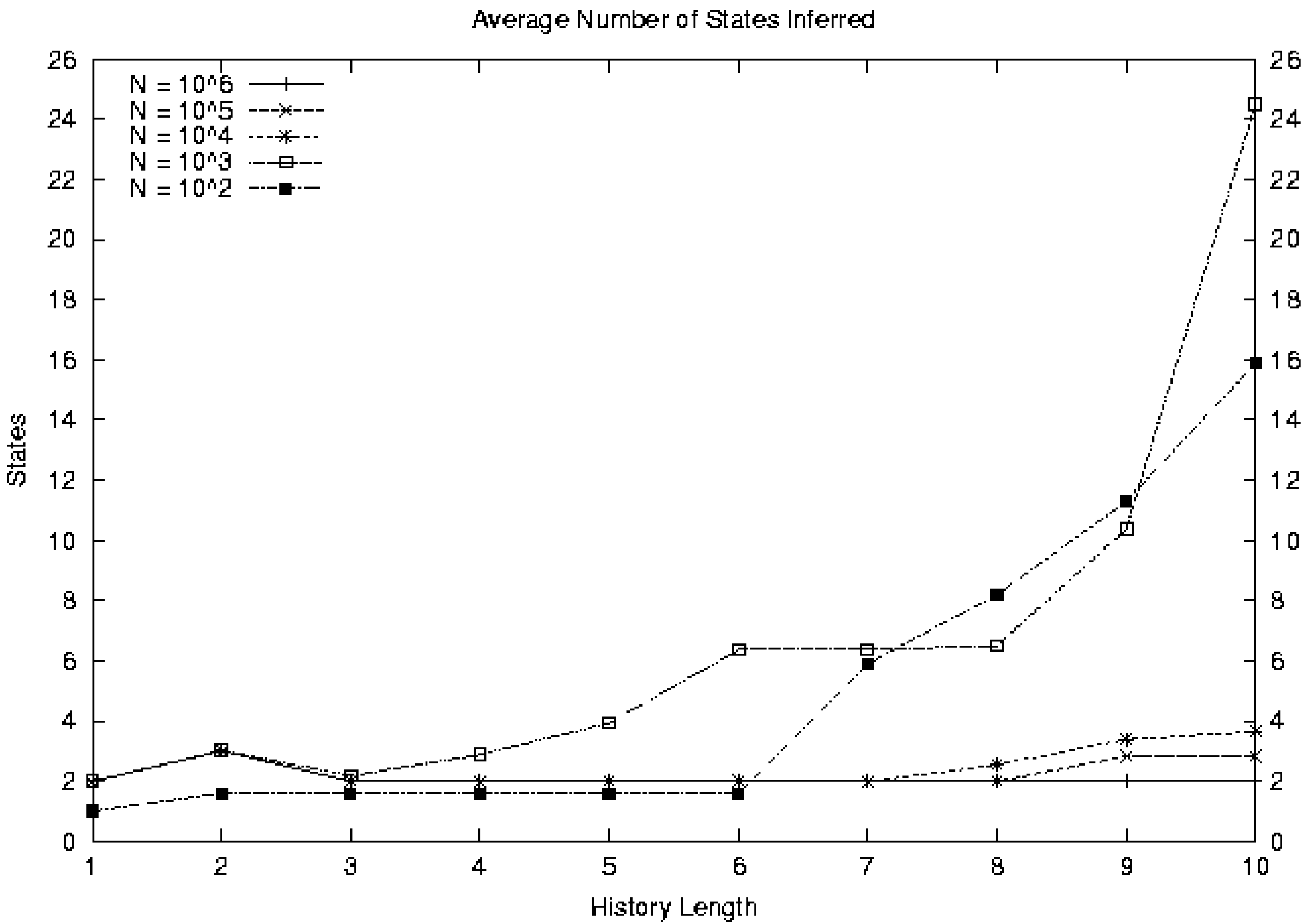
VLMM needs ∞ states, CSSR needs 2

Generally true of *sofic* processes



Prediction Error versus History Length
N = 10^6
N = 10^5
N = 10^4
N = 10^3
N = 10^2
Variational Error (log scale)
10
1
0.1
0.01
0.001
1
2
3
4
5
6
7
8
9
10
L


Scaled Error versus History Length
Rescaled Variational Error
L
N = 10^6
N = 10^5
N = 10^4
N = 10^3
N = 10^2
0
2
4
6
8
10
12
14
16
18
20
3
4
5
6
7
8
9
10


Average Number of States Inferred
N = 10^6
N = 10^5
N = 10^4
N = 10^3
N = 10^2
States
History Length
0
2
4
6
8
10
12
14
16
18
20
22
24
26
1
2
3
4
5
6
7
8
9
10

Results: Even Process

N	Distance CV	Distance CSSR	States CV	States CSSR
10^2	1.27 (0.23)	1.10 (0.23)	6.6 (1.5)	1.6 (1.0)
10^3	1.25 (0.41)	0.19 (0.23)	5.6 (1.7)	2.2 (0.1)
10^4	1.15 (0.02)	0.02 (0.02)	2.0 (0)	2.0 (0)

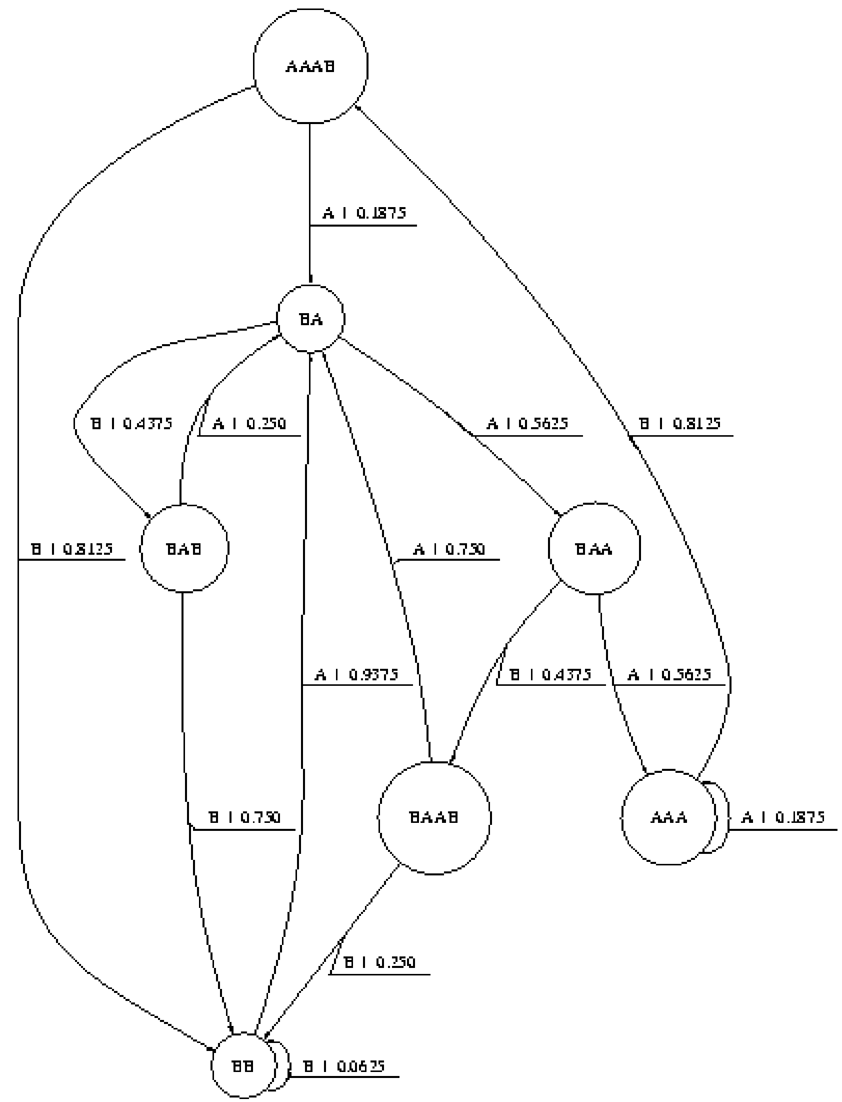
Foulkes process

7-state binary process

Introduced by Foulkes in 1959 paper (JANET)

Can be put in context-tree form

Used by Feldman & Hanna (1966) to study human learning



Results: Foulkes Process

N	Distance CV	Distance CSSR	States CV	States CSSR
10^2	1.41 (0.23)	0.70 (0.12)	4.5 (2.1)	5.1 (1.5)
10^3	1.40 (0.17)	0.21 (0.06)	5.8 (2.7)	6.6 (0.8)
10^4	1.40 (0.11)	0.06 (0.01)	2.3 (0.7)	7.2 (0.6)

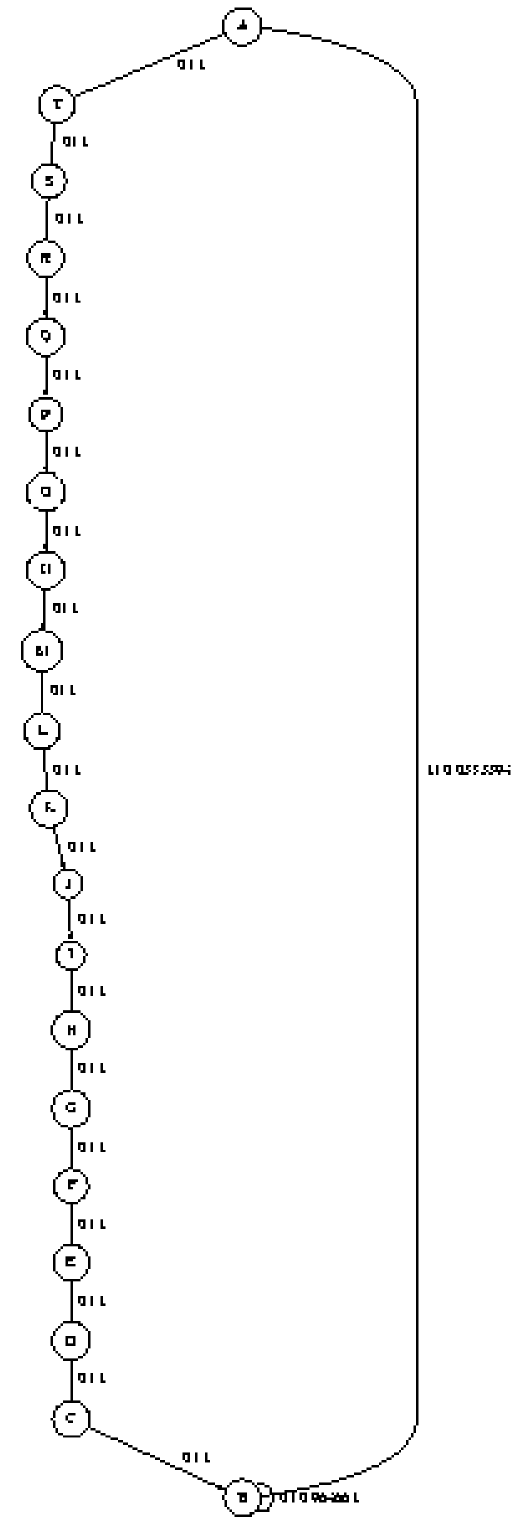
Model neuron

(Klinkner, Shalizi & Camperi, 2005)

One of a system of noisy neurons which synchronize each other

1 time-step = 1 ms

refractory period of 19 ms



Actual neuron

data courtesy G. Gage/UM Center for Neural Prosthetics

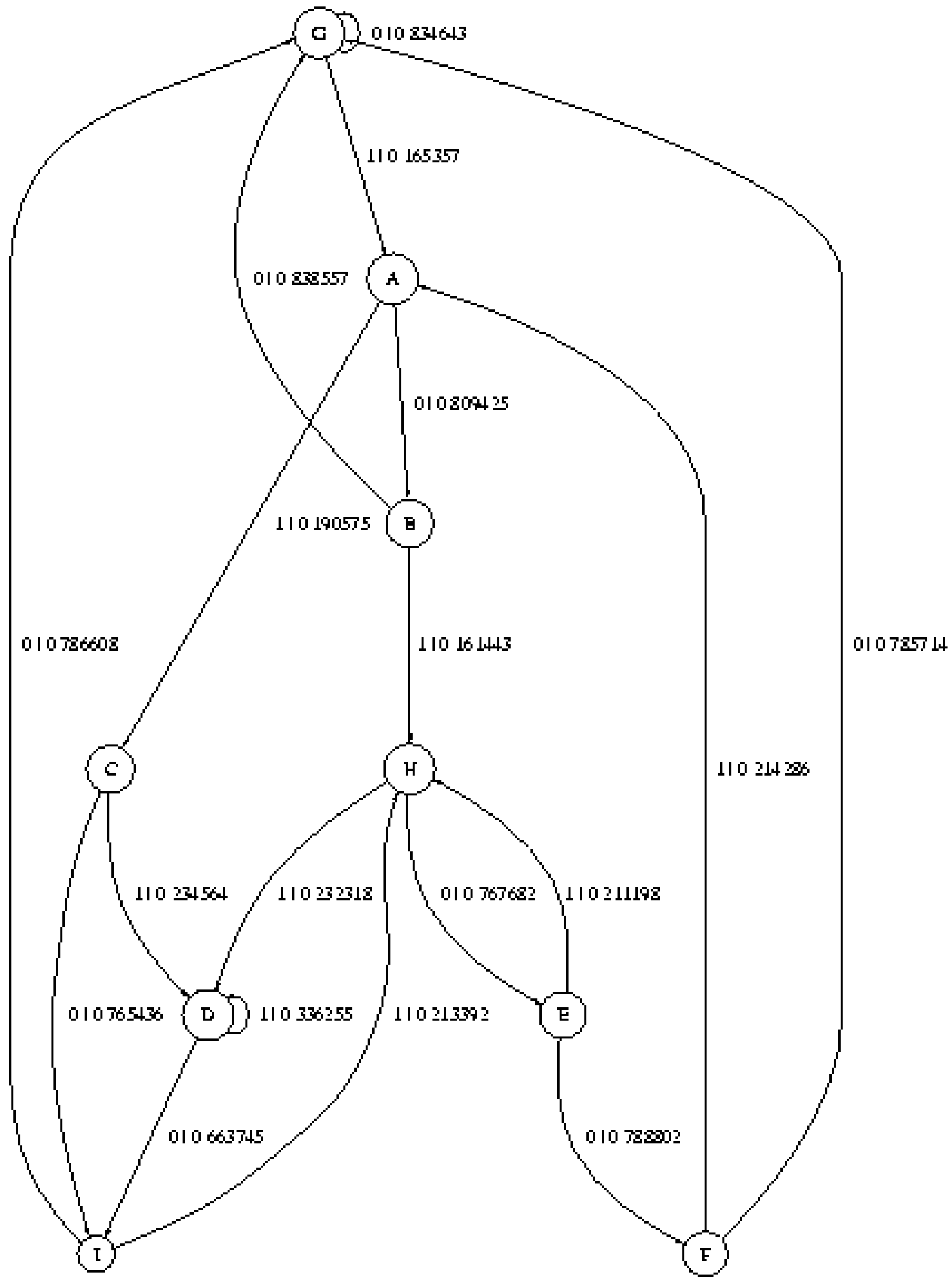
Multi-electrode array (“Michigan probe”) durably implanted in motor cortex of awake, behaving rat

Up to 16 units recorded simultaneously

Data from motor-learning experiment

This neuron: Quiescence, isolated spikes, bursts

Spike process, signal 34a
445
446
447
448
449
450
Time



G
010 834643
110 165357
010 838557
A
010 809425
110 190575
B
110 161443
010 786608
010 785714
C
H
110 214286
110 234564
110 232318
010 767682
110 211198
010 765436
D
110 336255
110 213392
E
010 663745
010 788802
I
F

Applications

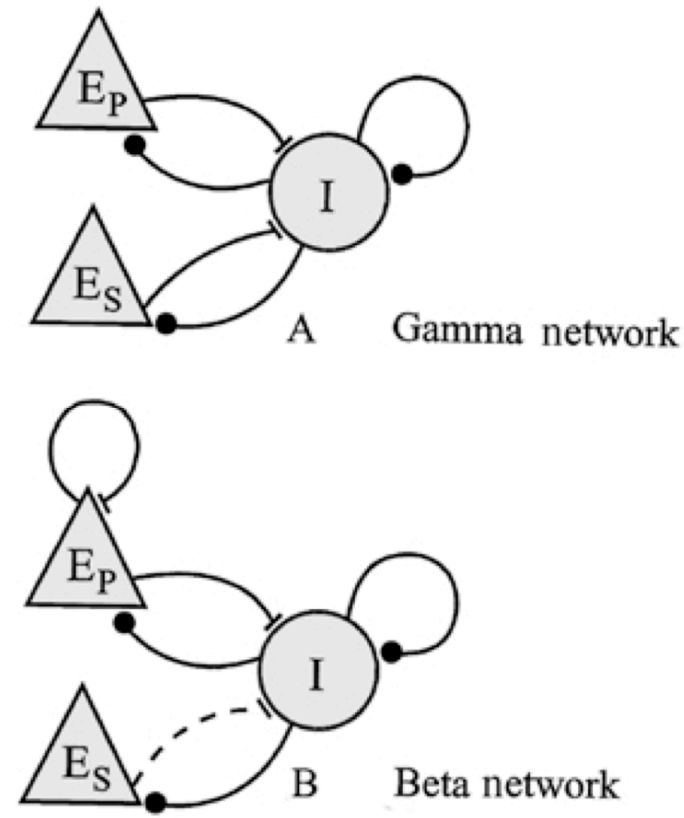
crystallography
  Varn and Crutchfield 2003
geomagnetic fluctuations
  Clarke, Freeman & Watkins 2003
anomaly detection
  A. Ray 2004
seismology
turbulent velocity series
natural language processing
  Padro, 2005, 2006
neural coherence
  Klinkner, Shalizi & Camperi 2005

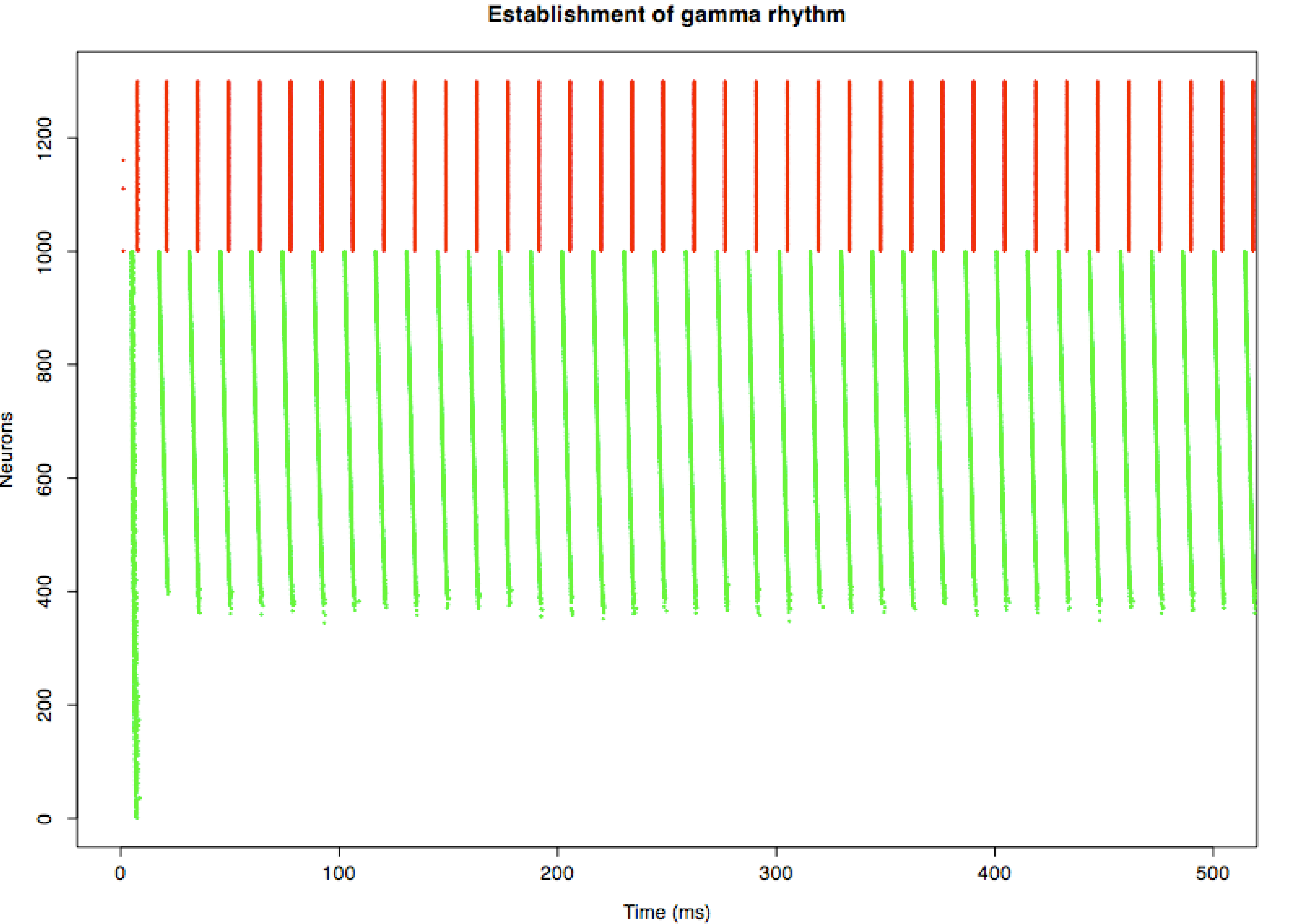
Application: Information-Sharing in Network Dynamics

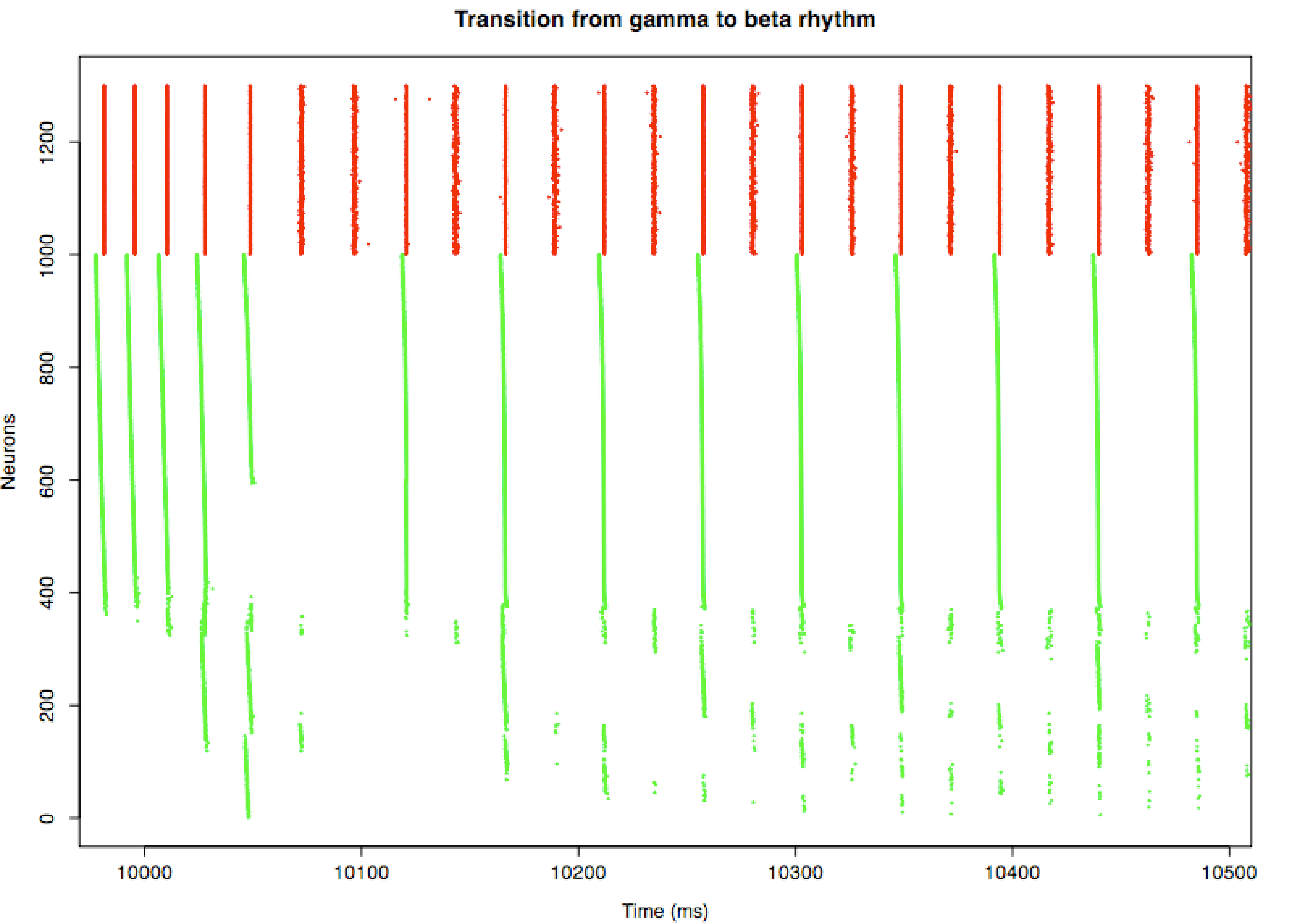
- Time series for each node
- Run CSSR on each node's time series
- Find series of states
  - Filters out un-predictive noise
- Calculate mutual information across states
- **Informational coherence** = normalized MI
- Positive IC => shared dynamical information
  - Much better estimates of information sharing than MI on original time series
  - More robust than other synchrony measures
- Can then find **functional communities**

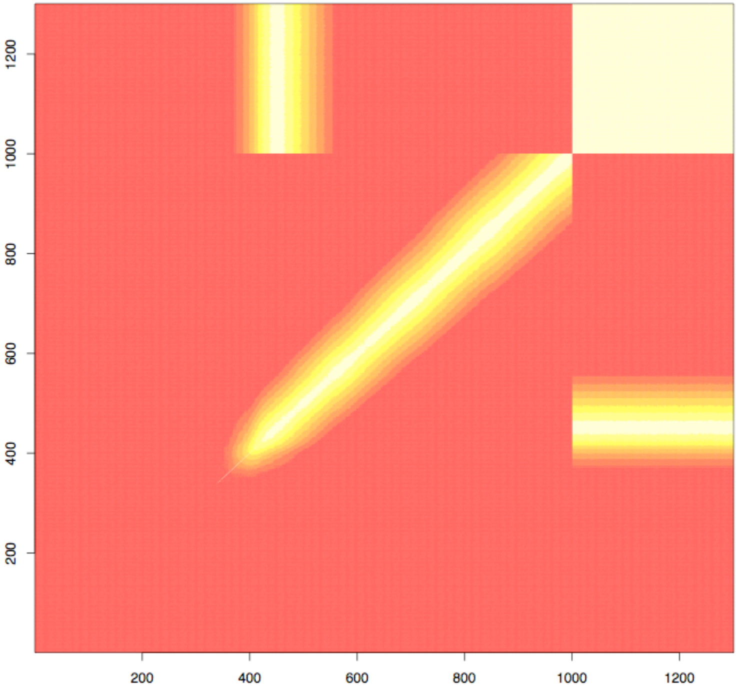
Test on Model Data

We use a network with two different neuron populations (1000 pyramidal cells and 300 interneuron), in which one of the populations (pyramidal cells) receive a heterogeneous drive, which divides the population into two classes: participating (P) and suppressed (S). The diagram shows the synaptic modifications that lead to the temporal separation between the two groups during beta. Through a Hebbian rule, recurrent excitatory connections are activated among the participating neurons, while connections among suppressed cells (dashed line) are weakened.

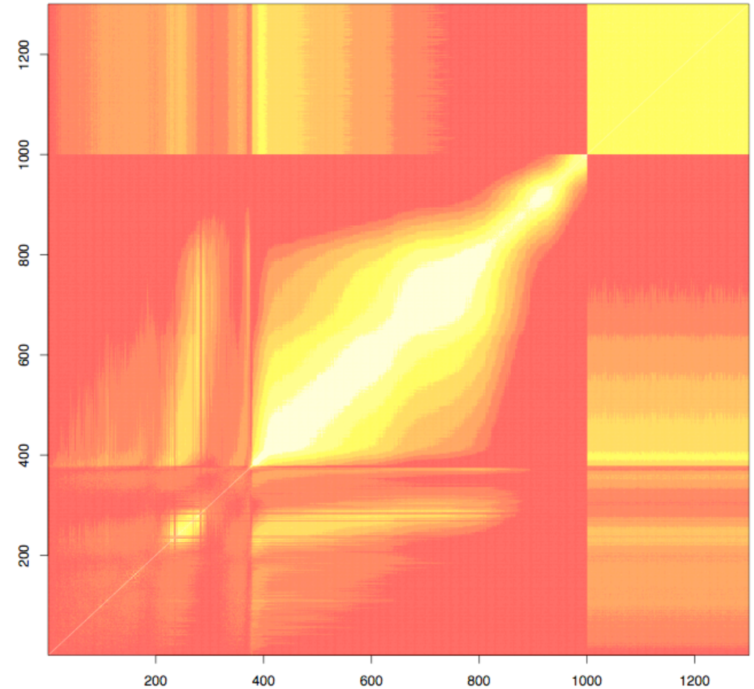



Establishment of gamma rhythm
Neurons
Time (ms)
0
200
400
600
800
1000
1200
0
100
200
300
400
500


Transition from gamma to beta rhythm
Neurons
0
200
400
600
800
1000
1200
10000
10100
10200
10300
10400
10500
Time (ms)

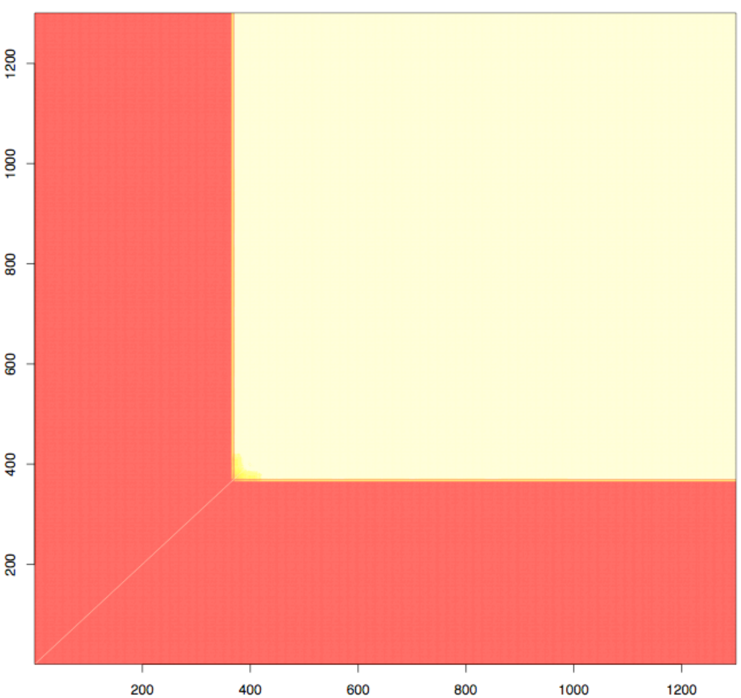


Cross-correlation

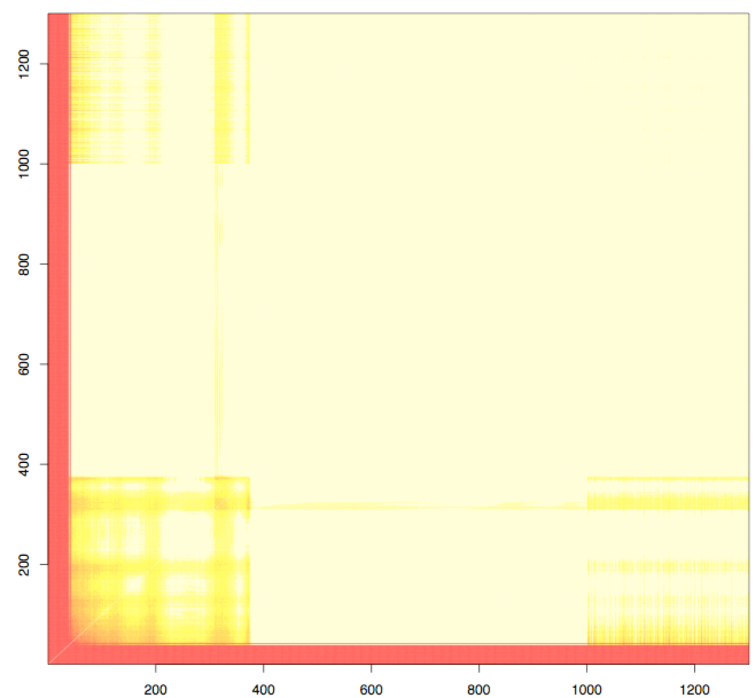


Gamma

Beta



Informational coherence



Extensions

Transducers, controlled dynamical systems

Spatio-temporal systems

Continuous-valued series ??

- Estimating generating partitions? (Kennel & Buhl, Hirata et al., Bollt et al., ...)
- Kernel density estimators?
- Adaptive discretization? (Boschetti unpub.)

Higher-order languages ????

Mostly self-organization

Spatio-temporal systems

Not-so-good ideas:

- one absurdly-dimensional time series
- many independent time series
- turning into a 1D system, pathwise

Better: Local statistics based on "light-cone"

Repeat the analysis to get local causal states and complexity

Self-organization

(Shalizi, Klinkner & Haslinger PRL 2004)

"I know it when I see it"

Disputes: turbulence, ecology,...

Does self-organizing $\Rightarrow$ irreversible?

- *Yes*: Priogine, Nicolis; Haken; etc.
- *No*: D'Souza, Margolus; Smith

Not self-organized criticality (necessarily)

Why not just use entropy?

- Low entropy disorganized systems (low-temperature stat. mech.)
- High entropy organized things (organisms)
- Organization ↑ *because* entropy ↑ (self-assembly)

System has self-organized between t_1 and t_2 if

(I) $C(t_1) < C(t_2)$

(II) the increase is not caused by outside input

Exorcism

Is the system being organized by its input?

Causal inference problem

Replace input with statistically-similar noise

  Delgado and Sole 1997

Exclude non-noise inputs

Cyclic cellular automata

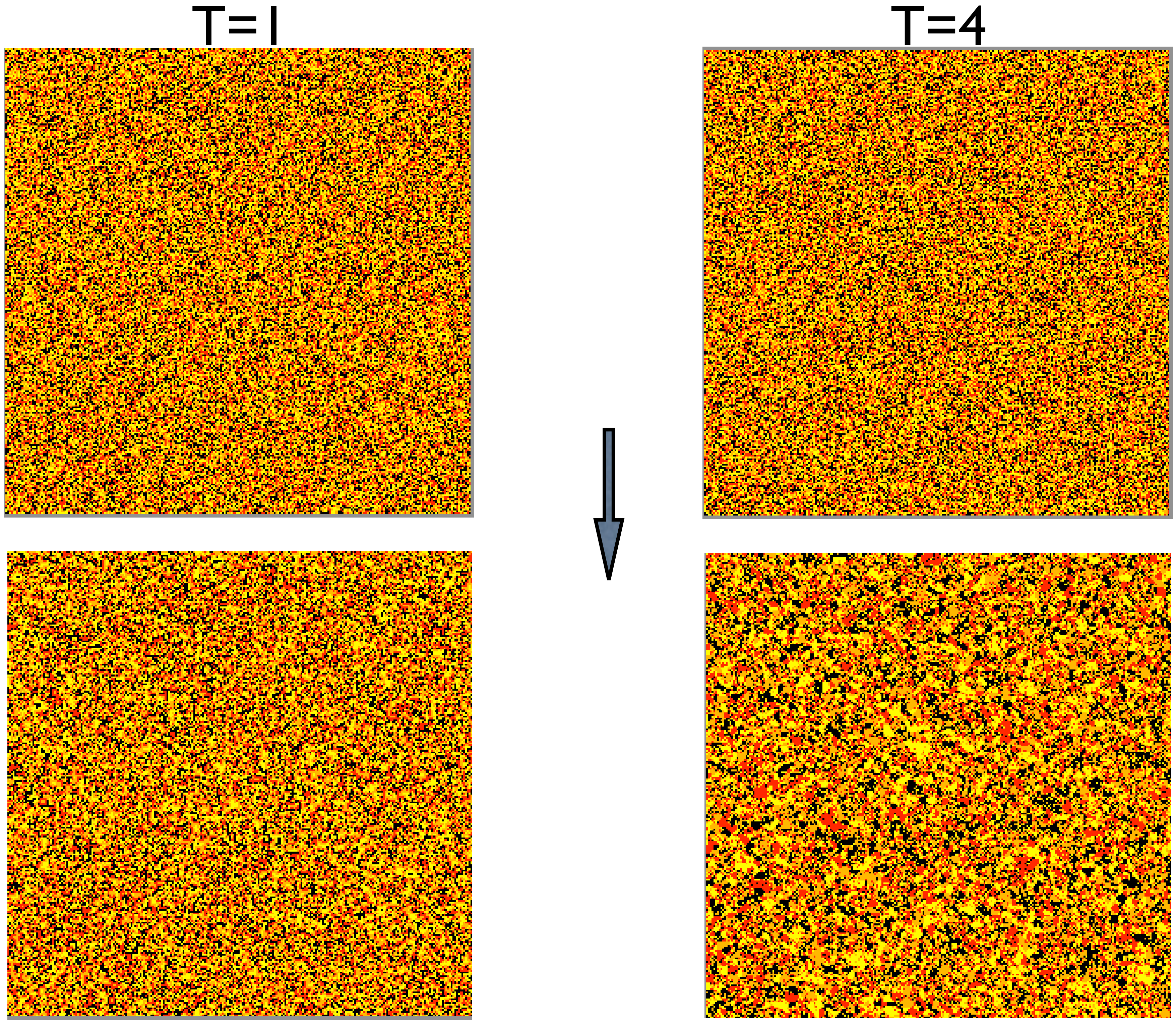
Qualitative model of excitable media

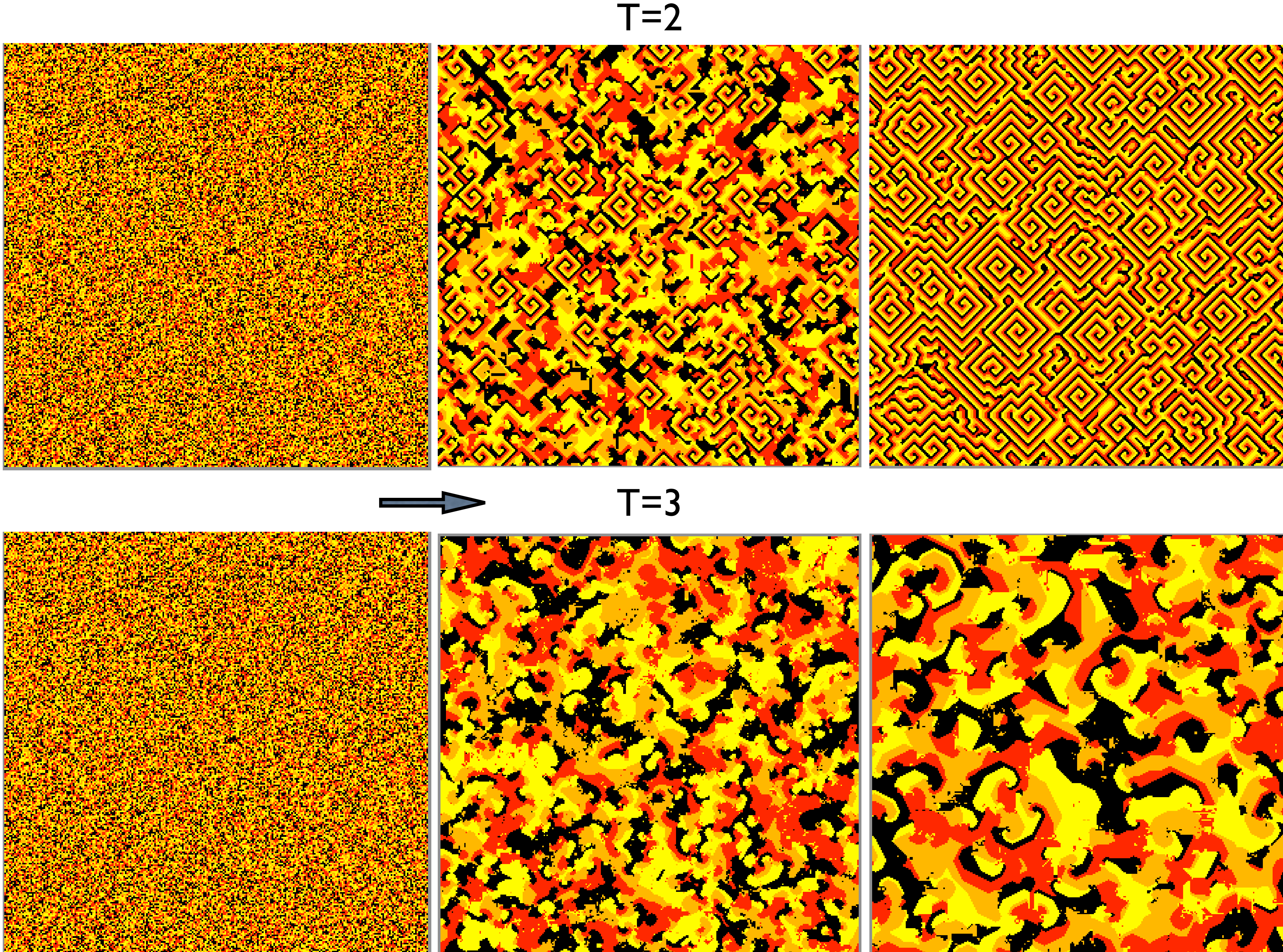
K colors; a cell of color k switches to k+1 (mod K) if at least T neighbors are already of that color

Analytic theory for structures formed

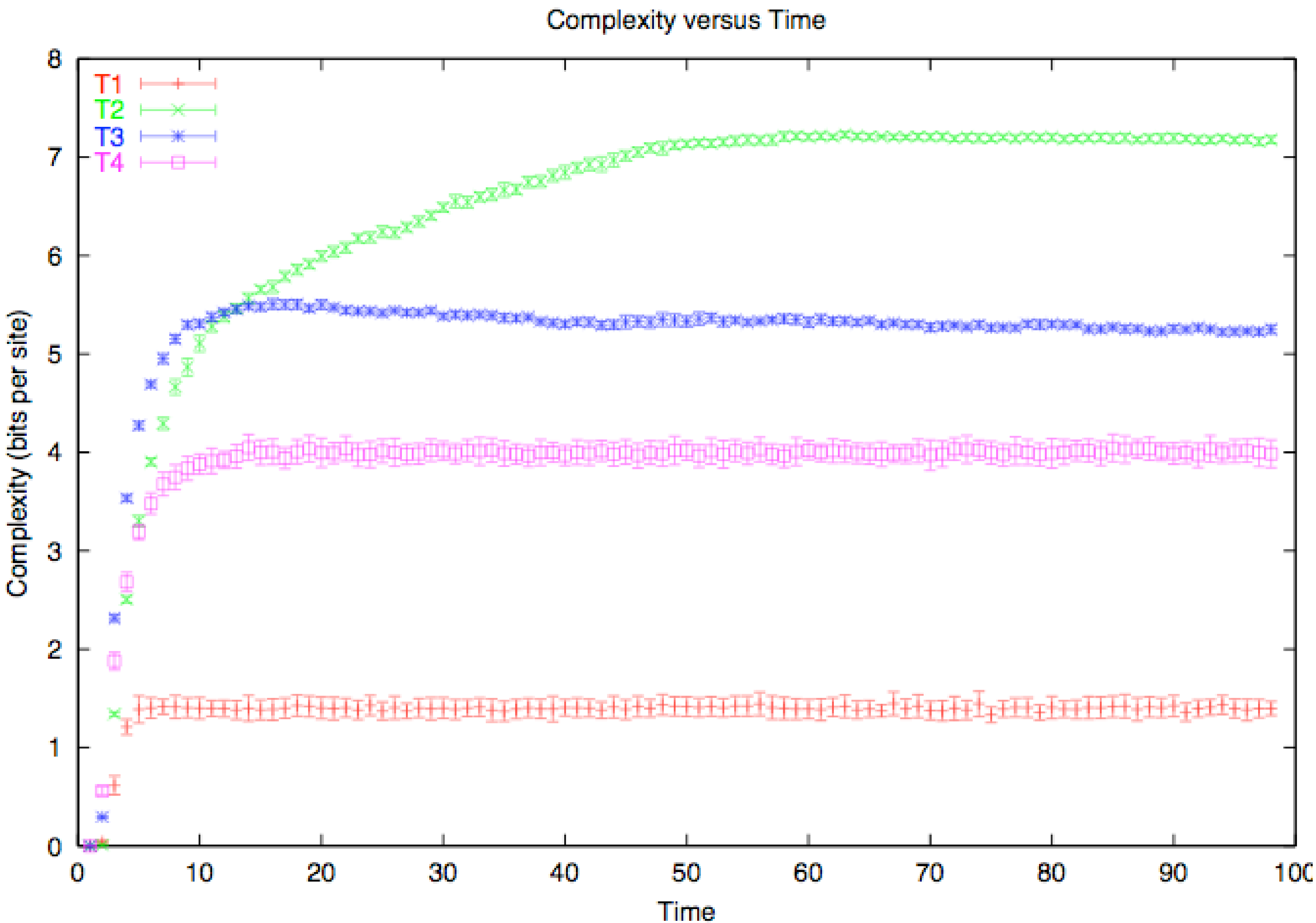
Griffeath et al.

Spirals, “turbulence”, local oscillation, fixation


T=1
T=4


T=2
T=3

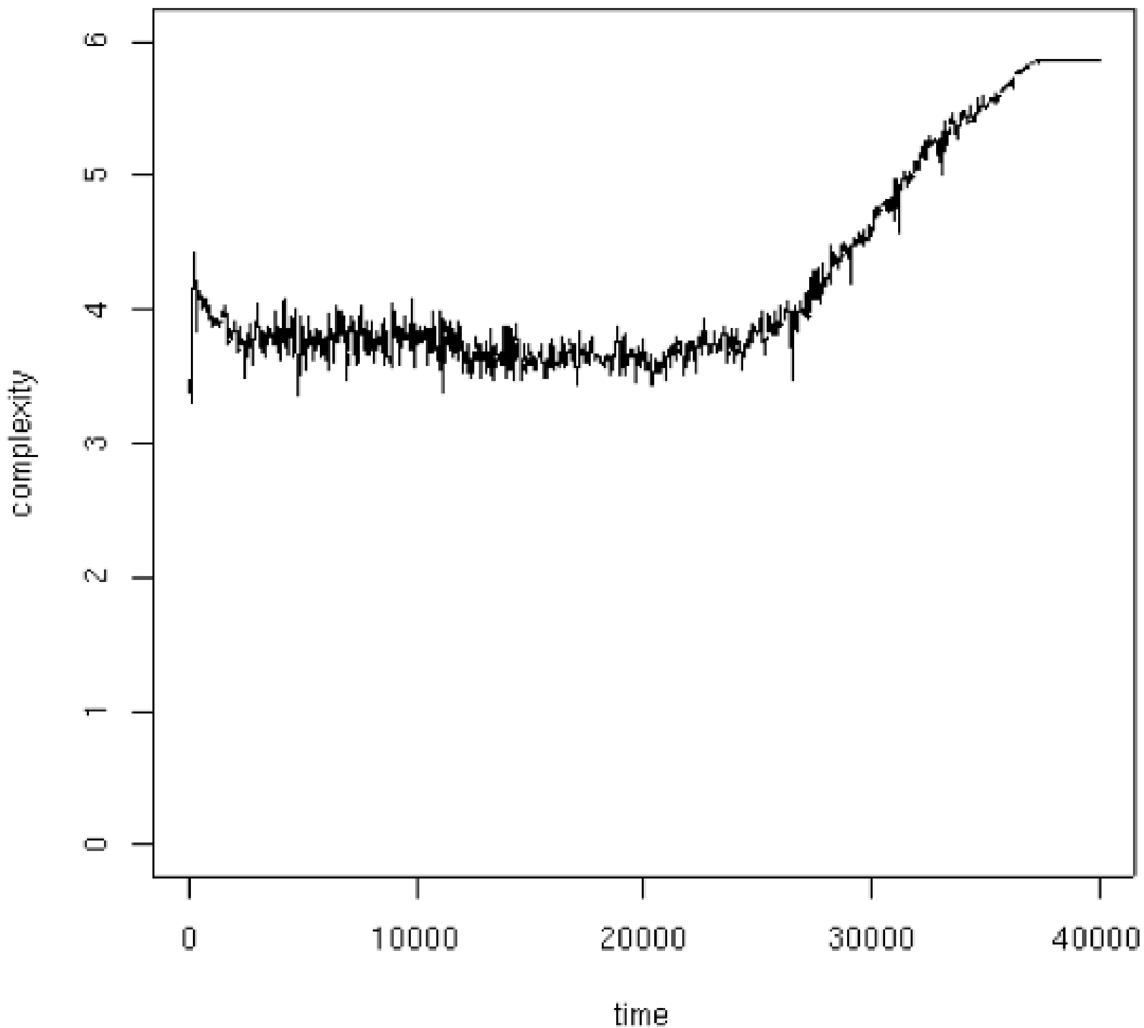
CCA



(300x300, n = 30)

BTW sand-pile

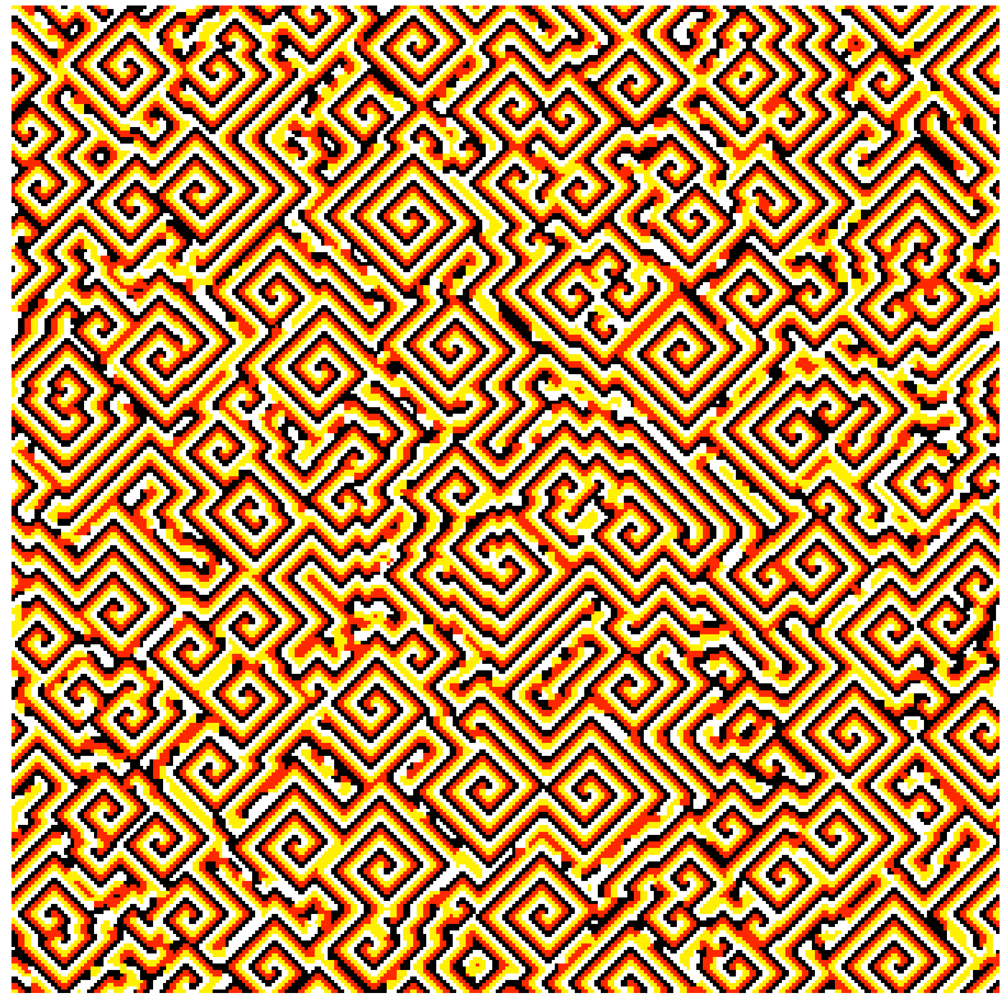
(J.-B. Rouquier, unpublished)



(supra-threshold relaxation, 300x300, n=1)

Finding Coherent Structures

(Shalizi, Haslinger, Rouquier, Klinkner & Moore, PRE 2006)



Spatially extended, temporally persistent

Generated by the micro dynamics

More efficient and more comprehensible descriptions ("emergent")

Order parameters

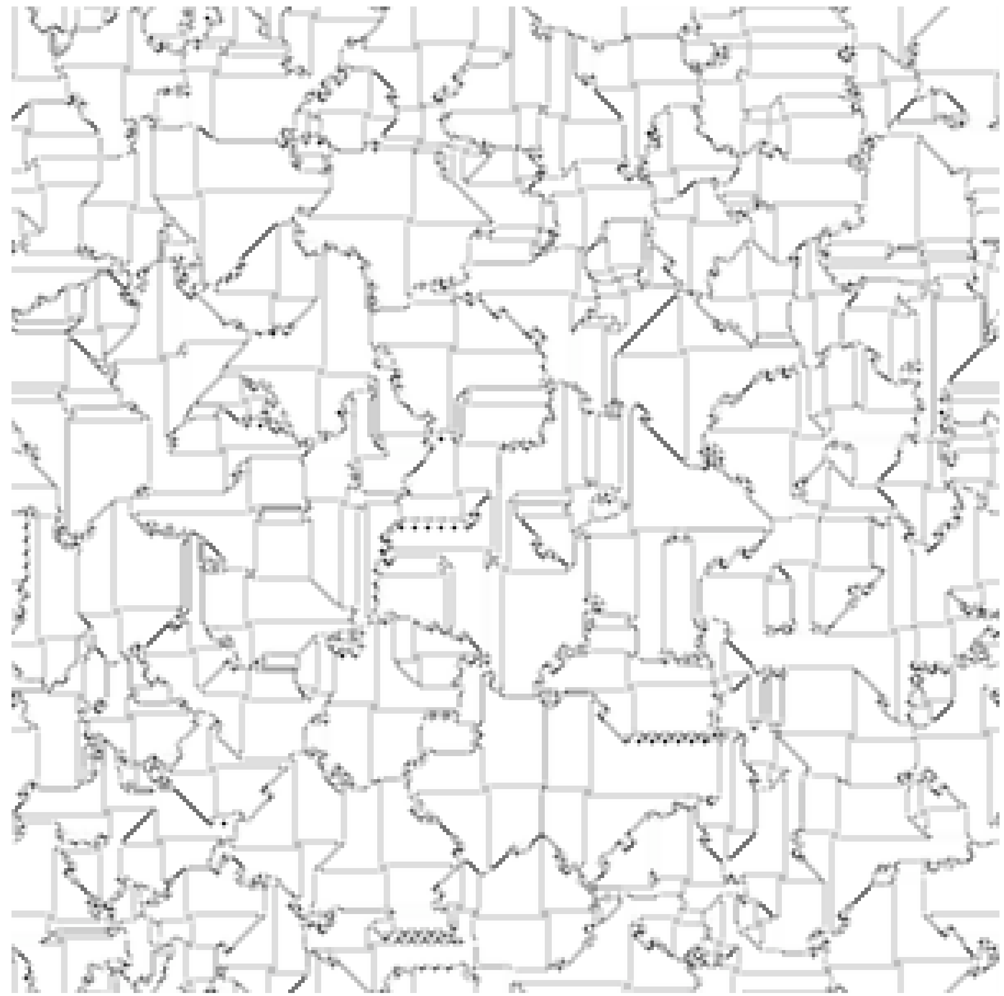
OP measures symmetry breaking

$$\Phi = f(OP)$$

$$-\log(\Pr(\text{config})) \propto \int \Phi \, dx$$

Structures = defects in OP field

OP found by trial and error

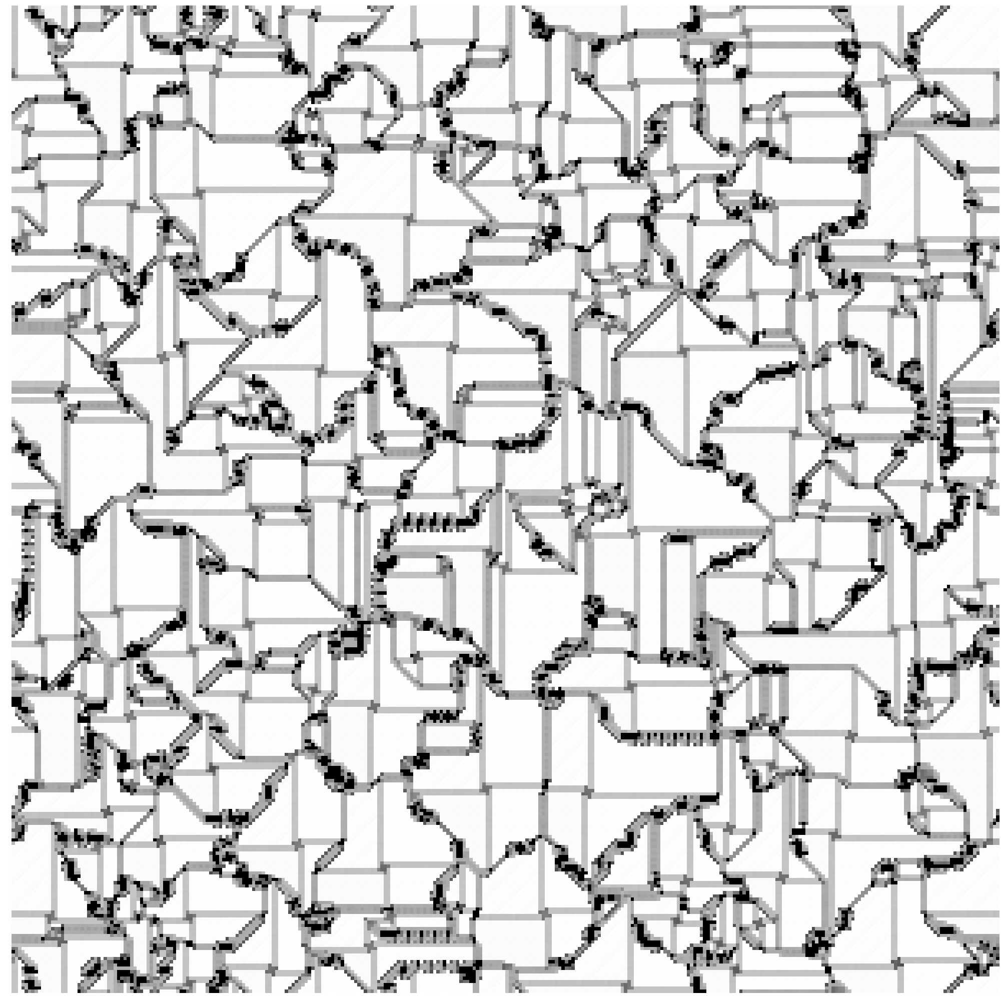


Complexity field

Local description length

$$C = -\log(\Pr(\text{state}))$$

Automatic; no tradition needed



Emergence

A logical relation between levels of description
The higher-level one is more interesting than the lower

- Thermodynamics emerges from statistical mechanics
- Chemistry from quantum mechanics
- Classical mechanics from quantum mechanics
- Superconductivity and Ohm's Law from quantum mechanics
- Demographic fluctuations from the 4Fs of animal behavior
- Evolutionary arms races from population genetics
- Efficiency (or bubbles) from microeconomic exchange
- Neurons, termites, ...

Ecosystems

Organisms

Organs

Functional systems

Tissues

Cells

Organelles

Metabolic networks

Macromolecules

Monomers

Atoms

Subatomic particles

(turtles↓)*

The bad idea

- “emergent” = “could not be predicted”
- predicted from what?
  - “water isn’t like hydrogen and oxygen”: so?
  - give us our interactions!
- predicted by who?
  - why should you care about my mathematical weakness?
- can computable systems show emergence?
- trivial or incomprehensible?
  - neither is fruitful

Try again

The higher levels are not as detailed
“Data abstraction”
Why hide details?
What do we want that information for?

Efficiency of prediction

(Palmer 2001)

Bits needed for prediction? $C = I[S_t;X^-_t]$

How many bits of prediction do you get?

Predictive information $E = I[X^+_t;X^-_t]$

Always need at least as much as you get

$$E \leq C$$

So efficiency is

$$0 \leq E/C \leq 1$$

For a Markov process

$$E = C - H[X_{t+1}|S_t]$$

Multiple levels

Low-level variables X
High-level variables Y, derived from X
Each has its own predictive structure
$\mathrm{eff}(X) \neq \mathrm{eff}(Y)$

A definition

If eff(Y) > eff(X), then Y *emerges* from X

Y abstracts *relevant* features from details of X

Depends on both the abstraction and the low-level dynamics

Different abstractions can emerge from the same low-level dynamics

- In the same situation (organs vs. functional systems)
- In different situations (Ohm's law vs. superconductivity)

Lattices, not chains or trees

Thermo

1 cc of argon at STP

At the molecular level, efficiency $\approx 10^{-9}$

(from scattering theory for entropy production)

At the thermodynamic level, efficiency ≈ 1

(from Onsager theory)

Gain of 10^9 $\therefore$ strongly emergent

Self-organization vs. emergence

Process over time on one level
vs. logical relationship between levels
Emergent properties if C(t) constant
- Thermodynamics, for instance

C(t) rising makes emergence more helpful
- "Why is my closet so full of my clothes?"

Extracting emergent variables

Nice, if we could do it!
All sorts of tricks for dimension reduction, feature selection, ...
Maybe: Look at the structure of the optimal predictor - it's already filtering for relevance

Summary

Predictive states are a powerful way of describing system complexity, and using it for prediction

Model discovery algorithms (especially CSSR) reliably reconstruct states from data

The theory can be applied to data

It also tells us something about self-organization and emergence

References

David Blackwell and M. A. Girshick (1954), *Theory of Games and Statistical Decisions* (Wiley)

James P. Crutchfield and Karl Young (1989), "Inferring Statistical Complexity", *Physical Review Letters* **63**: 105--108; http://www.santafe.edu/projects/CompMech/papers/ISCTitlePage.html

See also the other papers by Crutchfield and Young available from this site.

Cosma Rohilla Shalizi and James P. Crutchfield (2001), "Computational Mechanics: Pattern and Prediction, Structure and Simplicity", *Journal of Statistical Physics* **104**: 817--879; http://arxiv.org/abs/cond-mat/9907176

Predecessors and Successors

Wesley Salmon (1971), *Statistical Explanation and Statistical Relevance* (Pittsburgh: University of Pittsburgh Press)

Frank B. Knight (1975), "A Predictive View of Continuous Time Processes", *Annals of Probability* **3**: 573--596

Peter Grassberger (1986), "Toward a Quantitative Theory of Self-Generated Complexity", *International Journal of Theoretical Physics* **25**: 907--938

Herbert Jaeger (1999), "Observable Operator Models for Discrete Stochastic Time Series", *Neural Computation* **12**: 1371--1398; http://www.faculty.iu-bremen.de/hjaeger/pubs/oom_neco00.pdf

Michael L. Littman, Richard S. Sutton and Satinder Singh (2002), "Predictive Representations of State", *Advances in Neural Information Processing Systems* [NIPS] **14**: 1555--1561; http://www.eecs.umich.edu/~baveja/Papers/psr.pdf

Self-Organization and Local Causal States

Jordi Delgado and Ricard V. Sole (1997), "Collective-induced Computation", *Physical Review E* **55**: 2338-2344; http://www.santafe.edu/sfi/publications/wpabstract/199608070

Cosma Rohilla Shalizi (2003), "Optimal Nonlinear Prediction of Random Fields on Networks", *Discrete Mathematics and Theoretical Computer Science* **AB**: 11-30; http://arxiv.org/abs/math.PR/0305160

Cosma Rohilla Shalizi, Kristina Lisa Klinkner and Robert Haslinger (2004), "Quantifying Self-Organization with Optimal Predictors", *Physical Review Letters* **93**: 118701; http://arxiv.org/abs/nlin.AO/0409024

Cosma Rohilla Shalizi, Robert Haslinger, Jean-Baptiste Rouquier, Kristina Lisa Klinkner and Cristopher Moore, "Automatic Filters for the Detection of Coherent Structure in Spatiotemporal Systems", *Physical Review E* **73** (2006): 036104, http://arxiv.org/abs/nlin.CG/0508001

Emergence

A. Jay Palmer (2001), "An Efficiency Measure for Agent-Based Modeling", Swarmfest proceedings

Cosma Rohilla Shalizi and Cristopher Moore (2003), "What Is a Macrostate?", http://arxiv.org/abs/cond-mat/0303625

CSSR

Cosma Rohilla Shalizi and Kristina Lisa Klinkner (2004), "Blind Construction of Optimal Nonlinear Recursive Predictors for Discrete Sequences", *20th Conference on Uncertainty in Artificial Intelligence* (UAI 2004), pp. 504--511; http://arxiv.org/abs/cs.LG/0406011

Katalin Marton and Paul C. Shields 1994), "Entropy and the Consistent Estimation of Joint Distributions", *Annals of Probability* **23**: 960--977

Hidden Markov Models and Cross-Validation

Neil Gershenfeld, *The Nature of Mathematical Modeling* (Cambridge University Press, 1999)

Excellent reference on all manner of applied mathematical topics

Context Trees/VLMM

Dana Ron, Yoram Singer and Naftali Tishby (1996), "The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length", *Machine Learning* **25**: 117--149

Peter Buhlmann and Abraham J. Wyner (1999), "Variable Length Markov Chains", *Annals of Statistics* **27**: 480-513; http://www.stat.berkeley.edu/tech-reports/479.abstract1

Matthew B. Kennel and Alistair I. Mees, "Context-tree Modeling of Observed Symbolic Dynamics", *Physical Review E* **66**: 056209

J. D. Foulkes (1959), "A Class of Machines which Determine the Statistical Structure of a Sequence of Characters", *Wescon Convention Record* [Western Electronics Convention], vol. 4, pp. 66-73

Julian Feldman and Joe F. Hanna (1966), "The Structure of Responses to a Sequence of Binary Events", *Journal of Mathematical Psychology* **3**: 371--387

Applications of Causal States

Dowman P. Varn and James P. Crutchfield (2003), "From Finite to Infinite Range Order via Annealing: The Causal Architecture of Deformation Faulting in Annealed Close-Packed Crystals", *Physics Letters A* **324**: 299-307; http://arxiv.org/abs/cond-mat/0307296

Richard W. Clarke, Mervyn P. Freeman and Nicholas W. Watkins (2003), "Application of Computational Mechanics to the Analysis of Natural Data: An Example in Geomagnetism", *Physical Review E* **67**: 0126203; http://arxiv.org/abs/cond-mat/0110228

Asok Ray (2004)Approaching Sequential NLP Tasks with an Automata Acquisition Algorithm, "Symbolic dynamic analysis of complex systems for anomaly detection", *Signal Processing* **84**: 1114-1130

Muntsa Padro and Lluis Padro (2005a), "Applying a Finite Automata Acquisition Algorithm to Named Entity Recognition", in *Proceedings of 5th International Workshop on Finite-State Methods and Natural Language Processing*, http://www.lsi.upc.edu/~nlp/papers/2005/fsmnlp05-pp.pdf

---- (2005b), "A Named Entity Recognition System Based on a Finite Automata Acquisition Algorithm", Procesamiento del Lenguaje Natural, http://www.lsi.upc.edu/~nlp/papers/2005/sepln05-pp.pdf

---- (2005c), "Approaching Sequential NLP Tasks with an Automata Acquisition Algorithm", *Proceedings of International Conference on Recent Advances in NLP*, http://www.lsi.upc.edu/~nlp/papers/2005/ranlp05-pp.pdf